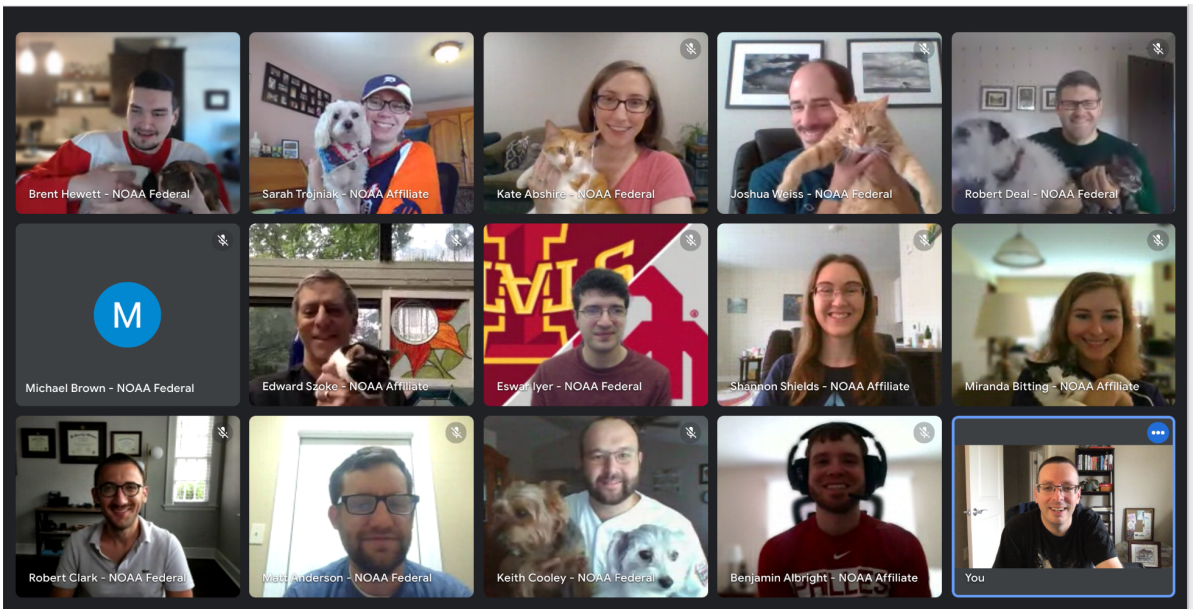
2021 Flash Flood and Intense Rainfall Experiment: *Findings and Results*



June 21 - July 23, 2021

Weather Prediction Center
Hydrometeorology Testbed

Sarah Trojniak - Systems Research Group, NOAA/NWS/WPC/HMT
James Correia Jr. - CIRES CU Boulder, NOAA/NWS/WPC/HMT

Table of Contents

1. Introduction

The 9th Annual Flash Flood and Intense Rainfall (FFaIR) Experiment was held virtually this year, spanning from June 21 to July 23, 2021. FFaIR is part of the Hydrometeorology Testbed (HMT) at the Weather Prediction Center (WPC). This was the second time FFaIR had to be in a virtual setting as a result of the ongoing pandemic. The FFaIR Experiment brings together researchers, forecasters, academia, and developers to evaluate, test, and use experimental guidance to aid in the prediction of heavy rainfall and flash flooding. It also is a catalyst for the conversations and collaboration among the various areas of expertise in the field, allowing for unique interactions between the different meteorology/hydrology fields that normally would not have the opportunity to interact.

2. Science and Operations

The basic design of FFaIR is a balance between using experimental guidance in a pseudo-operational setting, verifying that guidance, open and unfiltered discussion about relevant topics, and topical science seminars. Each week a new group of participants works together to help assess the experimental models and products. The four weeks of the 2021 FFaIR Experiment were:

Week 1: June 21 - 25
Week 2: June 28 - July 2
Week 3: July 12 - July 16
Week 4: July 19 - July 23

This year, FFaIR had 58 participants spanning across the four weeks of the experiment, including once again a forecaster from the German Meteorological Service (Deutscher Wetterdienst or DWD) as well as participants from the Environmental Modeling Center (EMC), the Global Systems Laboratory (GSL), and the Hazardous Weather Testbed (HWT). A full list of the participants and their affiliation can be found in Appendix A.

As the National Weather Service (NWS) works towards its goal of having a Unified Forecast System (UFS), with all their operational models centered around the Finite Volume Cubed-sphere (FV3) core, the number of experimental FV3 Convective Allowing Models (CAMs) has increased as EMC, GSL, and their partners work to determine the optimal configuration for a FV3-CAM. This resulted in a significant part of FFaIR focusing on evaluating different FV3-CAM configurations. In addition to these CAMs, an ensemble run in the cloud as part of the NOAA's Big Data Program[1] was evaluated along with several new Machine Learning (ML) Excessive Rainfall Outlook (ERO) products from Colorado State University (CSU). All summary of the guidance evaluated in FFaIR can be found in Table 1 and

[1] https://www.noaa.gov/organization/information-technology/big-data-program

a more in-depth description can be found in Appendix B or the 2021 FFaIR Operations Plan. All of the experimental goals and objectives are listed below:

- Evaluate the usefulness of operational and experimental products from high resolution convective-allowing deterministic and ensemble models for forecasting extreme rainfall and flash flood events, with the main focus on the Day 1.
- Collect more information on the prolific forecasting of grid point storms in the FV3-CAMS that were identified in the 2020 FFaIR Experiment (see the 2020 FFaIR Final Report). Identify aspects of the cells such as size, rain rate, timing, weather patterns, etc.
- Focus on the guidances' ability to correctly forecast precipitation events exceeding various thresholds such as 2, 4 and 6 inches.
- Evaluate models' and ensembles' timing of precipitation onset, progression, and end during a 6 h time period.
- Evaluate CSU MLP for the Day 1 ERO, which this year focuses on using CAM models for the forecasts.
- Explore using Average Recurrence Intervals (ARI) as the base for an excessive rainfall outlook. These will be referred to as an ARI-ERO.

Table 1: *Summary of the experimental guidance evaluated in the 2021 FFaIR Experiment.*

Summary of Models, Ensembles and Products for 2021 FFaIR	
EMC	Three versions of the FV3-Limited Area Model (LAM): FV3-LAM FV3-LAMX FV3-LAMDAX
EMC/GSL	Rapid Refresh Forecast System (RRFS) Ensemble run on the Cloud: RRFS-Cloud or RRFSCE
GSL	RRFS1
OU-CAPS	Storm Scale Ensemble Forecast (SSEF)
OU-CAPS	Four Members of the SSEF as deterministic models: SSEF Control Member RRFS-like Member HRRR-like Member WoFS-like Member
CSU	First Guess ERO Fields from: HRRR BLEND NSSL-sptavg -> NSSL2 NSSL-sptavg-landsea-mask -> NSSL3 NSSL-sptavg-landsea-mask-params -> NSSL4 NSSL-tempavg -> NSSL5

2.1 Daily Operations

Table 2 highlights the daily operations in FFaIR. Since the participants were spread across different time zones as a result of the virtual nature of the experiment, FFaIR started at 10 EDT (except on Mondays). The day began with a summary of the previous day's weather events and current conditions by the WPC forecaster before they jumped into the forecast discussion, which focused on output from the experimental models, ensembles, and products. The group then broke out into two different forecasting teams, one led by the WPC forecaster and the other by the FFaIR facilitator. Both groups created an ERO valid from 16 UTC to 12 UTC (valid the same period that the WPC Day 1 ERO is valid for) but the definition of the ERO for each group varied. The WPC forecaster led group created an ERO using the WPC definition of an ERO, identifying the risk of rainfall exceeding Flash Flood Guidance (FFG). The FFaIR facilitator led group created an ERO centered around exceeding rainfall Average Recurrence Intervals (ARIs). This product will hereafter be referred to as the ARI-ERO. Both these products will be explained in further detail in Section 2.2.To make sure within the first two days of FFaIR every participant had the opportunity to draw each type of ERO, Monday's teams were the same as Tuesday's just working on the opposite ERO. So if you were on the team that drew the ARI-ERO on Monday, on Tuesday you and your team drew the ERO. For the rest of the week, the teams were randomly assigned using the Google Meet Breakout Session randomizer. The creation of the EROs was usually followed by the verification session. This was done using Google Survey. The focus of the questions will be discussed in Section 2.3.

The afternoon was usually dominated by the second forecasting activity. This activity focused on using model data forecast precipitation amounts over a region of concern over a 6 hour time period. Both the region of interest and the valid forecast time were decided by the participants, though one requirement for the valid time was that the product could not be valid prior to 21 UTC. The product issued by each of the participants is referred to as the Maximum Rainfall and Timing Product (MRTP). It is a product developed by the FFaIR team, designed to mimic WPC's Mesoscale Forecast Discussion (MPD) while also trying to tease out information about models'/ensembles' tendencies; refer to Section 2.2 for more information on the product. In addition to the MRTP activity, on Tuesdays and Thursdays FFaIR Science Seminars were given. These seminars are composed of two invited speakers to present on FFaIR relevant research and products and are open to all of the NWS and its partners. A list of the seminars can be found in Appendix A.2.

Table 2: *The weekly schedule for 2021 FFaIR Experiment.*

Time (UTC)	Monday	Tuesday	Wednesday	Thursday	Friday
1300 – 1400		On your own - Situational Awareness	On your own - Situational Awareness	On your own - Situational Awareness	On your own - Situational Awareness
1330 – 1430	Call-in - Greetings and Orientation				
1400 – 1600	Call-in - Day 1 ERO Forecasting Activity	Call-in - Day 1 ERO Forecasting Activity	Call-in - Day 1 ERO Forecasting Activity	Call-in - Day 1 ERO Forecasting Activity	Call-in - Day 1 ERO Forecasting Activity
1600 – 1615	Break	Break	Break	Break	Break
1615 – 1745	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification
1745 – 1830	Lunch	Lunch	Lunch	Lunch	Lunch
1830 – 2030	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification	Call-in - Forecast Activity or Verification
1830 – 1930		DIFFERENT GOOGLE LINK - Science Seminar		DIFFERENT GOOGLE LINK - Science Seminar	

2.2 Forecast Activities

As stated, this year's forecasting activities were the ERO, ARI-ERO, and the MRTP[2]. Producing a collaborative Day 1 ERO has long been a staple in FFaIR. It parallels WPC's Operational ERO and is defined as "the probability that rainfall will exceed FFG within 40 kilometers (25 miles) of a point." The product is scheduled to be updated three times a day, with the last update at 16 UTC, resulting in a product that is valid from 16 UTC to 12 UTC. The risk categories are: Marginal (5-10%), Slight (10-20%), Moderate (20-50%), and High (>50%). The FFaIR ERO follows the same definition and valid time period as the last scheduled update for the Operational ERO. An example of an Operational and FFaIR ERO can be seen in Fig. 1A-B.

The creation of the FFaIR ERO is a collaborative effort, with participants actively discussing what they think will happen. Like last year, as an aid to help convey to the WPC forecaster drawing the collaborative ERO, participants used Google Slides to draw their own ERO on a static map. This allowed all participants, as well as the WPC forecaster, to quickly see similarities and differences in everyone's thoughts on what the ERO should look like. An example of a couple EROs drawn via Google Slides can be seen in Fig. 1C-D.

[2] Excessive Rainfall Outlook, Excessive Rainfall Outlook Average Recurrence Interval based, and the Maximum Rainfall and Timing Product.

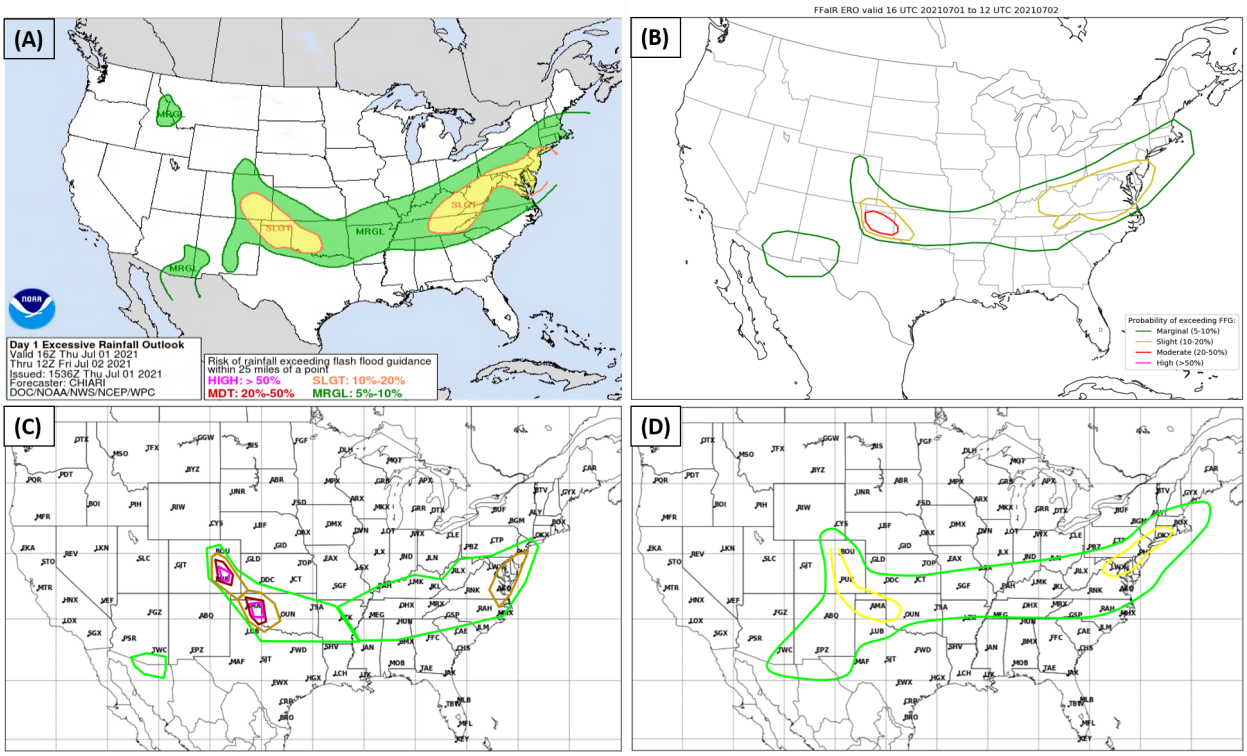


Figure 1: *The Day 1 (A) Operational ERO, (B) collaborative FFaIR ERO and (C)-(D) examples of the EROs drawn by the participants using Google Slides, valid 16 UTC 01 July to 12 UTC 02 July, 2020. Marginal - green, slight - yellow, moderate - red, and high - pink.*

The ARI-ERO was a new forecast activity in FFaIR this year. It was created in an attempt to begin to identify the association of flash flooding and ARI exceedances. Since this was the initial goal, six hour ARI exceedances were chosen rather than 24 hour exceedances even though the ARI-ERO was a Day 1 product. The six hour exceedance was chosen because often it is a lot of rain in a short time that leads to flash flooding, so the FFaIR team felt using a shorter time frame for the ARI exceedance would be a better proxy to capture the flash flood reports. Therefore the product identities were in any given six hour time interval within the valid time of the product (16 UTC to 12 UTC) rainfall has a 75% chance of exceeding the 1, 2, 5, and 10 year ARI. Examples of this product from each week can be found in Fig. 2.

Although the goal of this product and what the product highlights differs from the ERO, creating it followed the same methodology. The team would work together, discussing the forecast and how they felt the day would play out. Then using Google Slides they would draw what they thought the ARI-ERO should look like. Using this and advice from the participants, the FFaIR facilitator worked to draw the final, collaborative ARI-ERO product.

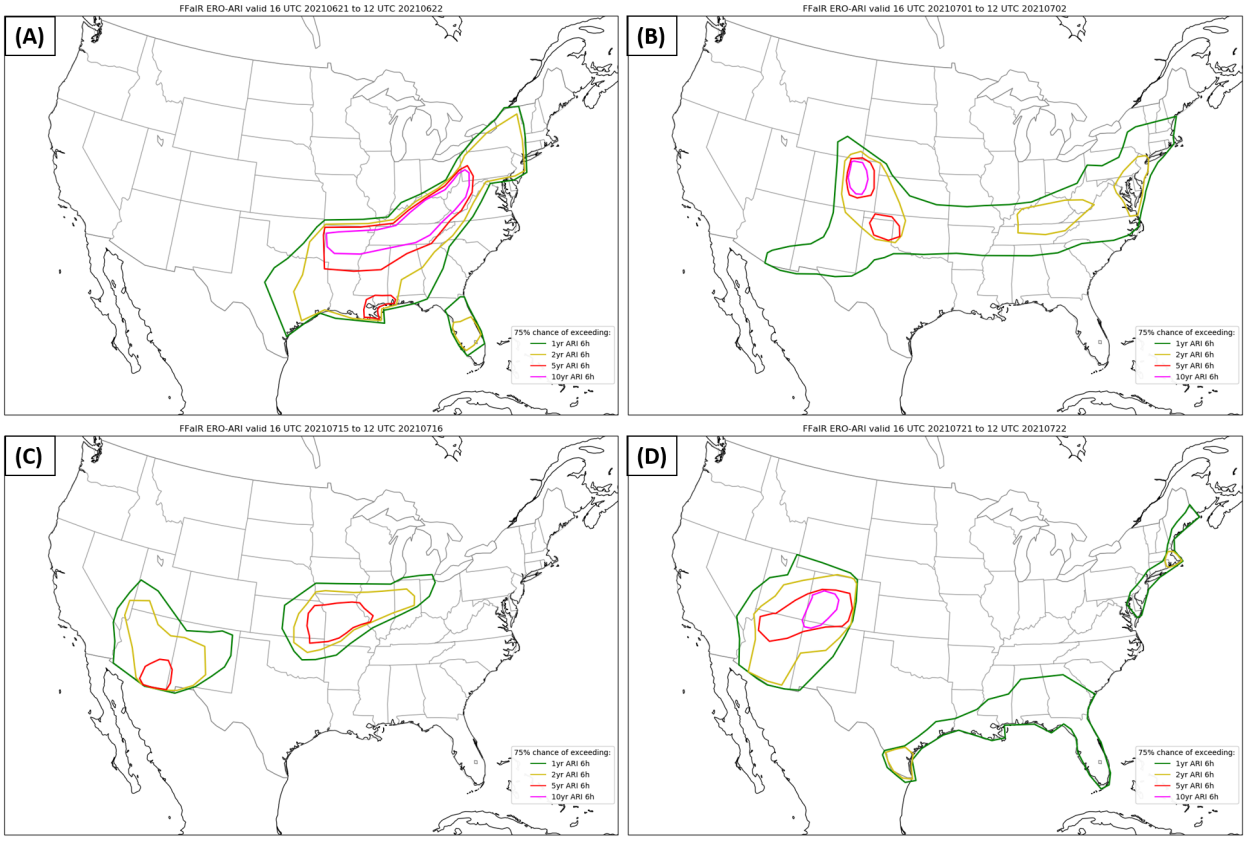


Figure 2: *The Day 1 ARI-ERO valid (A) 16 UTC 21 June to 12 UTC 22 June 2021, (B) 16 UTC 01 July to 12 UTC 02 July 2021, (C) 16 UTC 15 July to 12 UTC 16 July 2021, and (D) 16 UTC 21 July to 12 UTC 22 July 2021. 75% chance of exceeding the 6 h: 1y ARI - green, 2y ARI - yellow, 5y ARI - red, 10y ARI - pink.*

With the success of the MRTP last year, the forecast activity for the near-/short-term QPF product was once again the MRTP. The product is a combination of drawing contours for various thresholds and completing a survey while the forecast is being made. Each day the participants would pick a region and a time for the product. They then were randomly assigned a model or ensemble to evaluate while creating their MRTP, however they were not required to use it. Instead, in the survey, they had to state whether they did use their assigned model in their forecast and why or why not. The survey that the participants had to fill out as part of the MRTP can be found in Appendix C. Examples of MRTPs issued on July 01, 2021 can be seen in Fig. 3. Like last year, the participants had the option to draw contours for six hour rainfall totals of 1 inch, 2 inches, 3 inches, and 4 inches and to identify where they thought the highest rainfall total would occur. New this year, they also could highlight where they thought rainfall rates would exceed 1in./h.

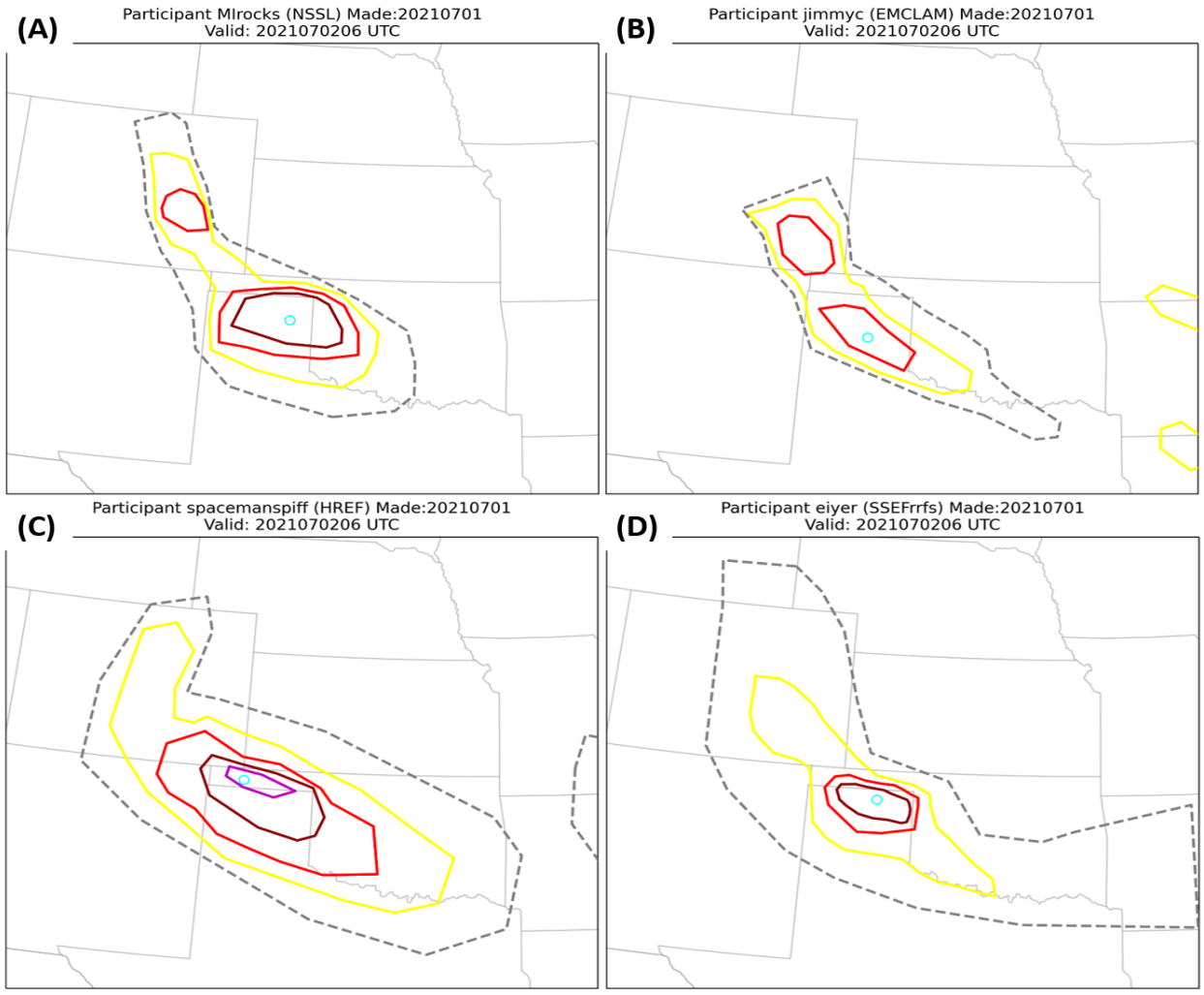


Figure 3: *Example of four MRTPs issued on July 01, valid from 00 UTC to 06 UTC July 02, 2021. Usernames for the MRTPs: (A) MIrocks, (B) jimmyc, (C) spacemanspiff, and (D) eiyer. Contours: yellow - 1 in., red - 2 in., dark red 3 in., purple - 4 in., and dashed gray - 1in/h rainfall rates. The blue circle is the forecasted location of maximum rainfall.*

2.3 Verification Methods

Verification of the guidance evaluated during FFaIR and the forecast products issued was done subjectively by the participants and then objectively after the completion of FFaIR. Subjective verification refers to verification that does not use statistical methods to evaluate the "goodness" of the guidance but rather how good humans felt the forecast was. This can also be thought of as a qualitative analysis while objective verification would be quantitative. Verification generally fell into the following groups:

1. Deterministic Quantitative Precipitation Forecast (QPF)
2. Deterministic Instantaneous Precipitation Rates (hereafter p-rate)
3. Ensemble QPF means and probabilities

4. CSU “first guess” EROs, FFaIR ERO and ARI-ERO
5. MRTPs

Multi-Radar Multi-Sensor Gauge Corrected (hereafter MRMS) precipitation data was used for verification of QPF. This was used rather than STAGE-IV precipitation data since the latency for the MRMS data is less than that of STAGE-IV and allowed for verification of the forecast the following day. To allow for consistency, MRMS was then also used in the objective analysis. However, FFG and ARI exceedances used to evaluate all the ERO products used STAGE-IV data as it is generally available by 16 UTC and the verification images for ERO products can be plotted quickly, thus the latency of the STAGE-IV was not an issue. Additionally, STAGE-IV is the data source that is currently used for verification of the WPC ERO, thus using it allows for consistency in ERO verification.

2.3.1 Subjective Verification

Deterministic 24 h QPF was evaluated by the participants for both the 00z and 12z model runs when available, valid 12 UTC to 12 UTC. This was done by showing the models’ 24 h QPF alongside the 24 h MRMS Quantitative Precipitation Estimate (QPE) and asking the participants to score the models’ forecast across the CONUS[3] on a scale of one to ten, with one being the worst and ten being the best. The scores could only be in whole numbers. During this question, the participants had access to an image like the one seen in Fig. 4 as well as access to Method for Object-Based Diagnostic Evaluation (MODE) information; this type of verification is discussed further in Section 2.3.2.

Deterministic QPF was evaluated in a timing sense as well. By timing we mean things like models being too slow/fast with convective initiation or progressing a system too slow/fast compared to observations. The goal of this question was to try and identify if an “incorrect” QPF forecast was because the model got the totals wrong or if the model actually got the rainfall totals right but just had them occurring too soon/late. For instance, if a model missed the QPF maximum during the 6 hr time period, was it because the model was incorrect on the amount of rainfall in 6 h or was it because the timing of the correct amount was slightly off? The idea to try and evaluate timing in the model stems from comments made by previous years’ participants when trying to evaluate 6 h QPF on the topic, as well as forecaster saying that sometimes they look at the 6 h QPF forecasts around the 6 h time period of interest to get a general sense of what could happen rather than focusing on the specific 6 h time period forecast from the model.

For the timing analysis, participants were assigned one of four models (HRRR, RRFS1, LAMDAX or the SSEF CNTL member). They were then asked to examine the 6 h window and domain for the previous day’s MRTP forecast and answer the question seen in Fig. 5. Participants were able to look at a static image, like the one in Fig. 6, as well as use a scroll bar to look at the 6 h QPF at and around the valid time period and compare it to the MRMS analysis

[3] CONtinental United States.

to determine if a timing issue impacted the QPF forecast at the valid time. For example, verifying for 00-06 UTC 2 July 2021, the MRTP domain was over the Central and Southern Plains (see Fig. 3). Looking at Fig. 6, the two middle images are the 6 h QPF and the MRMS, Fig. 6C and 6D respectively, valid at the 06 UTC (aka our time of interest). Participants were directed to first compare these differences between the model forecast and observed rainfall, focusing both on amounts and areal footprint of the 1" contour. Then, using the same image, they were told to look at the top/bottom two images which show the 6 h QPF valid one hour before/after (Fig. 6B/Fig. 6E) and two hours before/after (Fig. 6A/Fig. 6F), to see if a 6 h rainfall forecast before/after the valid time more closely resembled the MRMS analysis. If one of the time periods did, that might suggest there was a timing issue in the model. The participants then used the slider feature to further investigate how timing might have resulted in discrepancies in the model forecast and observations.

Model p-rate was verified against the p-rate from MRMS. Note that these two p-rates are not 100% comparable. The models have timesteps varying from 20s to 36s versus the MRMS which is every 2 minutes. The evaluation was framed as a comparison question, asking the participants to compare an experimental model's p-rate to both the HRRR's p-rate and the MRMS p-rate. Figure 7 shows the questions asked and Fig. 8 shows an example of what the verification images looked like. Like with the timing question, the participants were assigned which model they were to verify:

- 00z HRRR v LAMDAX
- 00z HRRR v RRFS1[4]
- 00z HRRR v SSEF CNTL member[5]
- 12z HRRR v LAMDAX
- 12z HRRR v RRFS1

P-rate was evaluated for the six hours that the MRTP was valid, so the participants were required to look and click through each hour's comparison.The setup of the verification image, which again can be seen in Fig. 8, was always HRRR p-rate on the left, MRMS p-rate in the center and the experimental model's p-rate on the right. The maximum p-rate was also plotted on the image from each of the datasets.

[4] Sometimes referred to as RRFS dev 1
[5] SSEF - Storm Scale Ensemble Forecast. This is also referred to as the OU CNTL member or CAPS CNTL member.

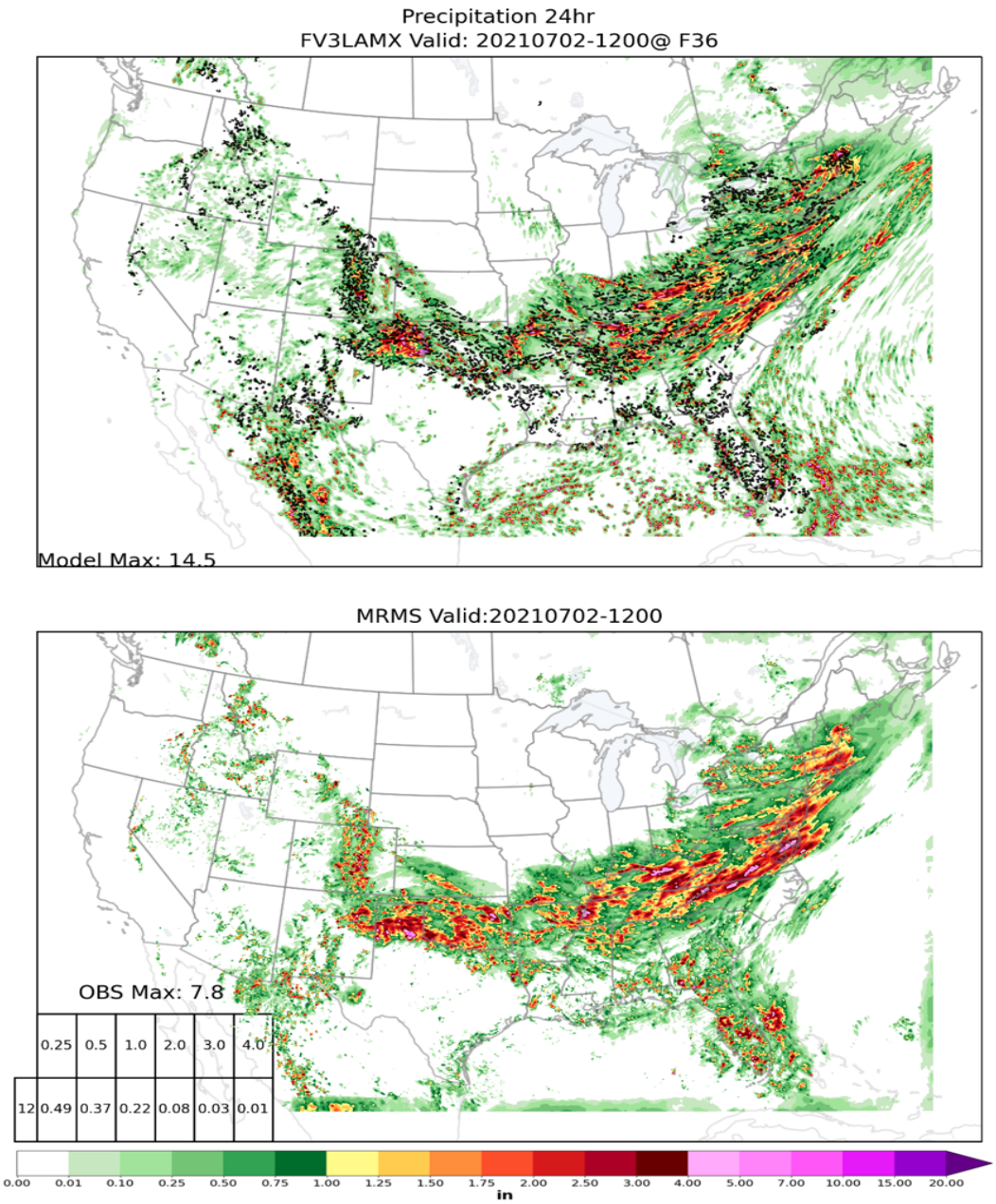


Figure 4: *Example of the 24 h QPF verification image valid 12 UTC 01 July to 12 UTC 02 July, 2021; (top) 00z LAMX and (bottom) MRMS. Plotted along with the QPF on the top image in dashed black is the observed area of 1+ inches of rain. Both images include their maximum rainfall amount. The bottom image also includes the CSI[6] at thresholds: 0.25, 0.5, 1, 2, 3, and 4 inches.*

[6] Critical Success Index

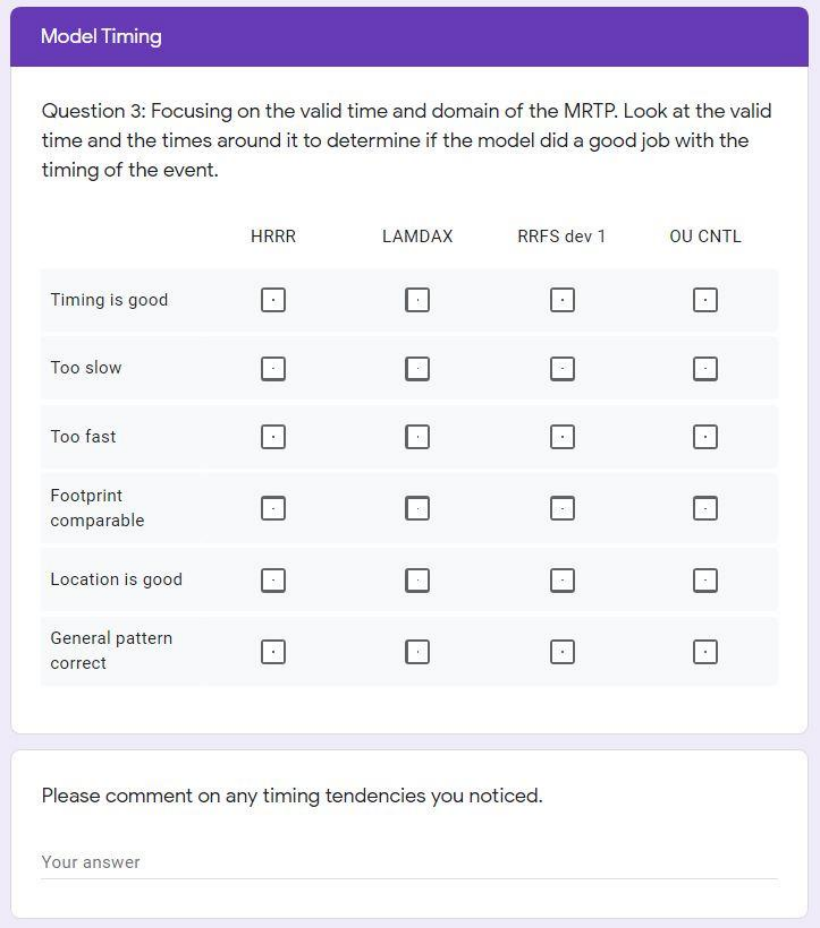


Figure 5: *The question asked and response format for the Deterministic QPF timing question.*

In previous years, subjective verification for the ensembles was done by looking at mean products. To try and gather information on how forecasters use the other information from ensembles, the subjective evaluation this year revolved around three hour and six hour exceedance probabilities. Focusing on the 00 UTC to 06 UTC time frame, participants were shown the 2 in/6 h probabilities, the 1 in/1 h probabilities for 00-03 UTC and 03-06 UTC, and the MRMS valid at 06z. An example of what they were shown for each ensemble can be seen in Fig. 9. They were instructed to look at the CONUS rather than a specific region of the domain. They then were asked, "1" (2") in 3hrs (6hrs) corresponds in some way to the observations. How?" The choices they could pick were: too high, high, about right, low, too low. They were also prompted to describe "how the probabilities do or do not correspond with that event." All models were initialized at 00z.

Additionally, since a goal for the new RRFS system is to extend both the deterministic runs and the ensemble runs out to 60 h, participants were asked to evaluate the Day 2 forecast from the RRFSCE and SSEF to begin to quantify the utility of CAM ensembles in the extended range. The Day 2 analysis was for the same valid time as the Day 1 (i.e. the 00 - 06 UTC time period), except the ensemble was initialized 24 hours earlier. The Day 2 question was set up the

same way as Day 1 but did not include verification of the HREF since it does not extend out to the valid forecast time.

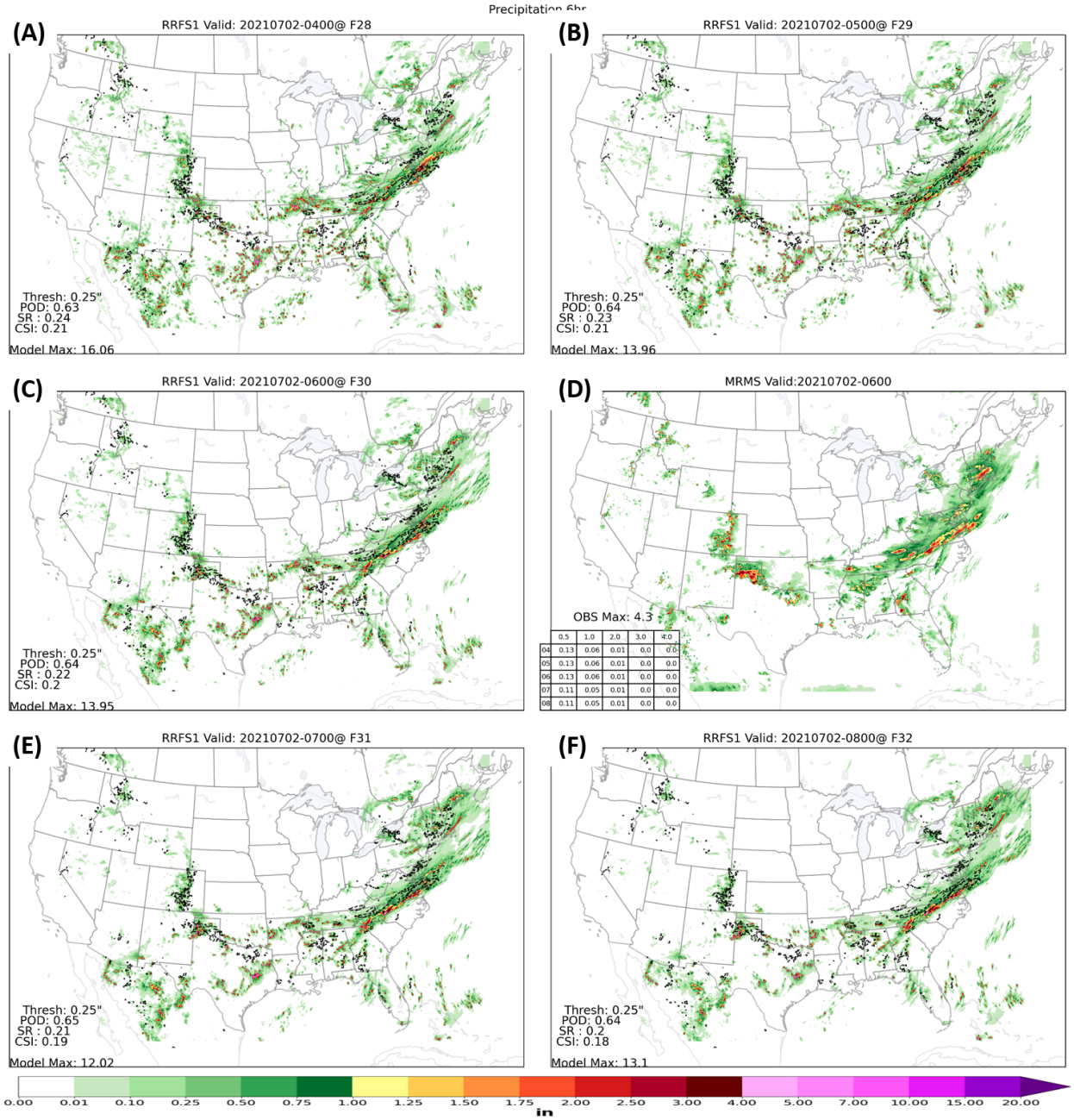


Figure 6: *Example of the static image for the 6 h QPF timing question, verification for the 6 h QPF at 06 UTC 02 July for the RRFS1. (C) QPF and (D) MRMS valid 00 UTC to 06 UTC. (A) and (B) are the 6 h QPF valid 2 h (22 UTC to 04 UTC) and 1 h (23 UTC to 05 UTC) prior to the verification time respectively. (E) and (F) are the 6 h QPF valid 2 h (01 UTC to 07 UTC) and 1 h (02 UTC to 08 UTC) after the verification time respectively.*

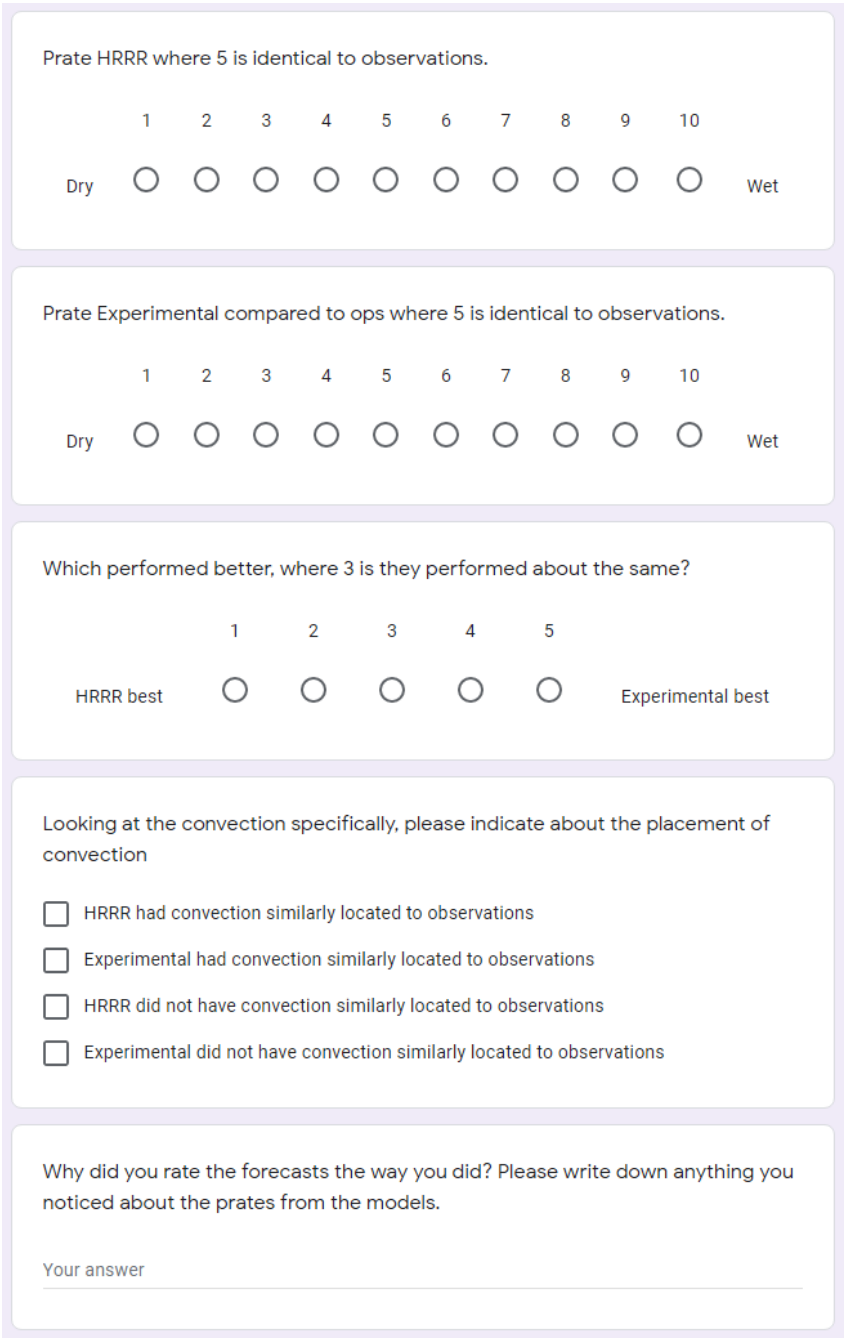


Figure 7: *The questions asked and response format for the deterministic p-rate verification.*

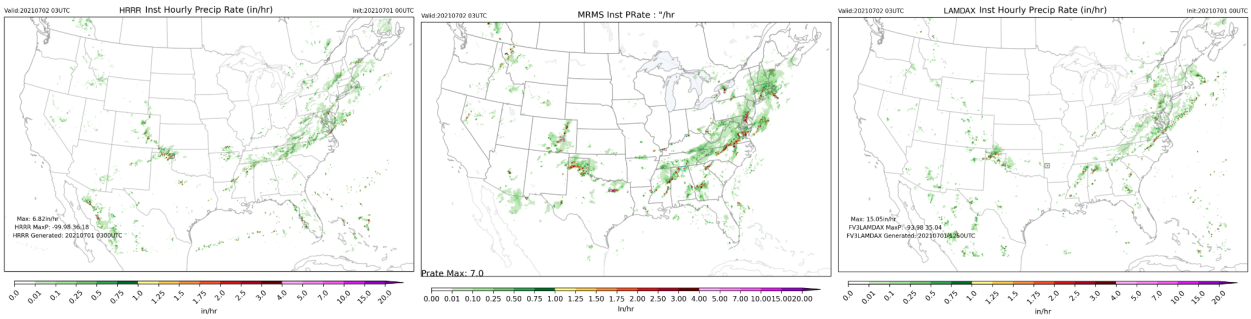


Figure 8: *Example of the set up for p-rate comparison. (Left) HRRR, (middle) MRMS, (right) RRFS1, valid 03 UTC 02 July 2021.*

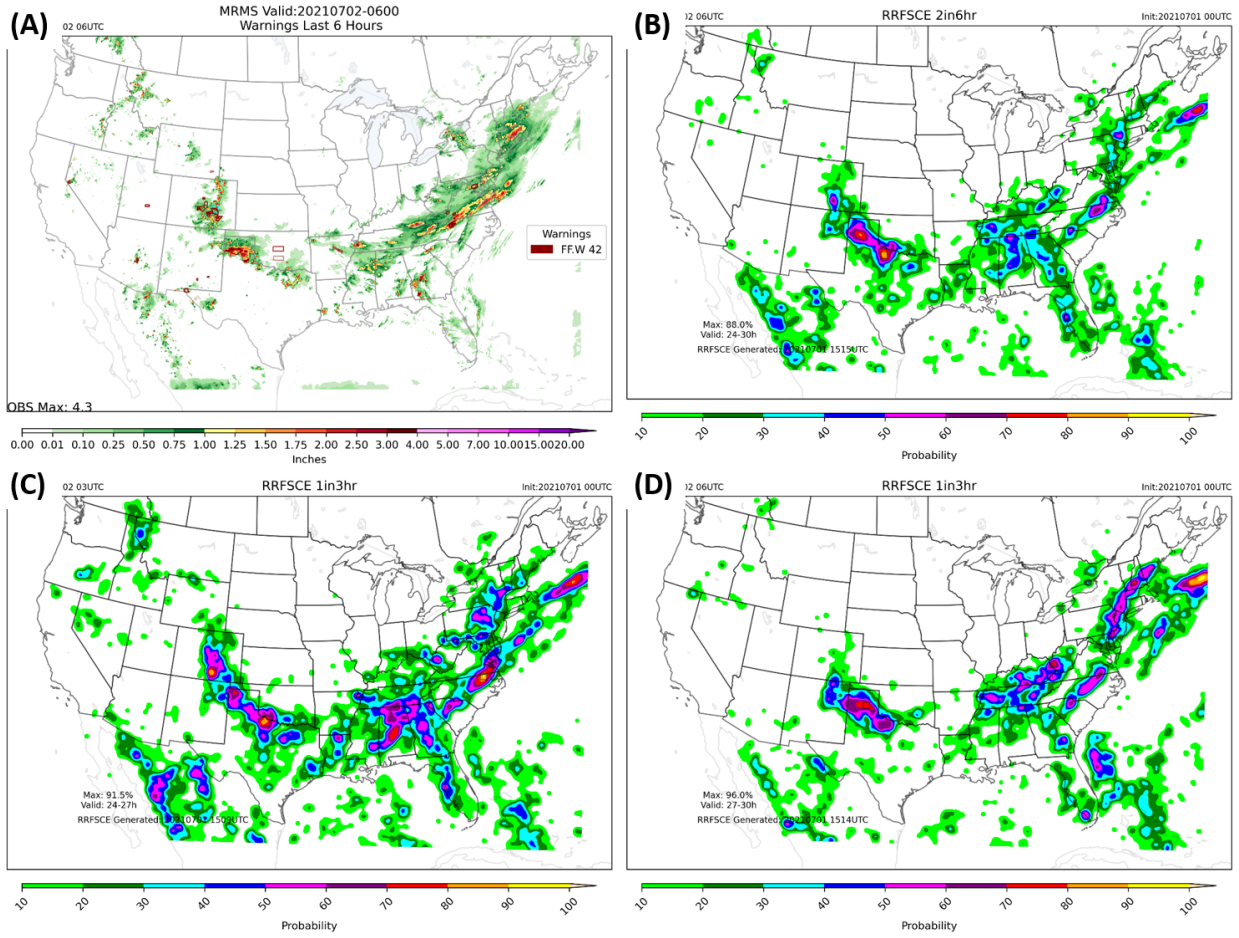


Figure 9: *Example of the set up for evaluation of ensembles. (A) the 6 h MRMS and (B) the RRFSCE probability of exceeding 2in/6h valid 00 UTC to 06 UTC 02 July 2021. The RRFSCE probability of exceeding 1in/3h valid (C) 00 UTC to 03 UTC and (D) 03 UTC to 06 UTC 02 July 2021.*

Subjective verification for the CSU "first guess" EROs, the FFaIR ERO and the ARI-ERO were all rated on their "goodness" on a scale of 1 (poor) to 10 (great). To help the participants determine this, the CSU and FFaIR EROs were plotted along with point observations from the Unified Verification Forecast System (UVFS), the MRMS valid the same time the ERO

was valid and the Practically Perfect ERO verification product. Both the Practically Perfect ERO verification and the UVFS will be discussed in further detail below. An example of what the participants saw for the verification of the EROs can be seen in Fig. 10. For the verification of the ARI-ERO, participants were shown the 6 h ARI exceedances of 1, 2, 5, 10, and 25 years plotted on top of the ARI-ERO. They were also shown the MRMS valid for the same time period, and the NOAA Atlas ARI images for the 6 h 1 y ARI and the 6 h 10 y ARI. An example of what the participants saw for the verification of the ARI-ERO can be seen in Fig. 11.

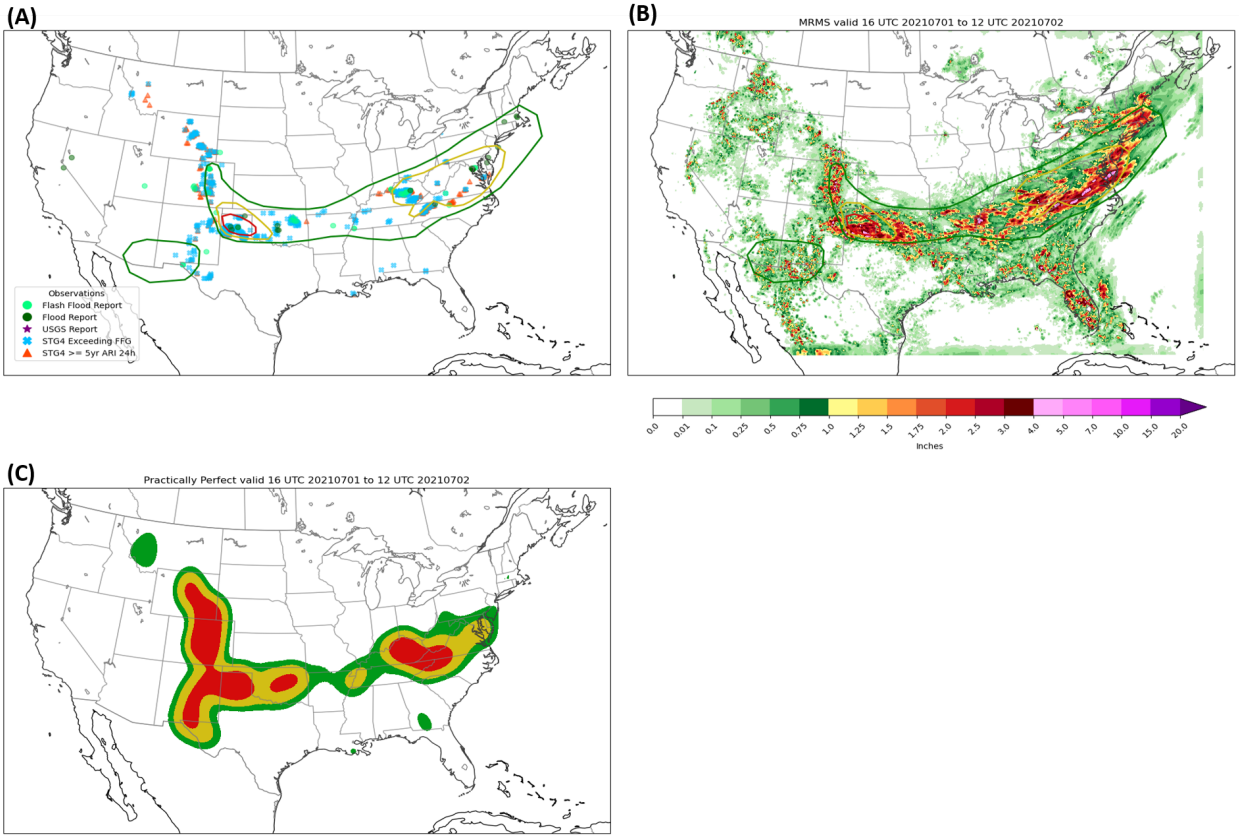


Figure 10: *Example of a subjective verification image for the CSU MLP First Guess forecasts and FFaIR Day 1 ERO. (A) FFaIR Day 1 ERO overlaid with: QPE > FFG (blue x), QPE >5 y ARI (orange triangle), flood/flash flood LSRs (light/dark green dots), and USGS gauge reports (purple star). (B) 20 h MRMS-GC QPE and (C)Practically Perfect analysis. Valid 16 UTC 01 July to 12 UTC 02 July, 2020.*

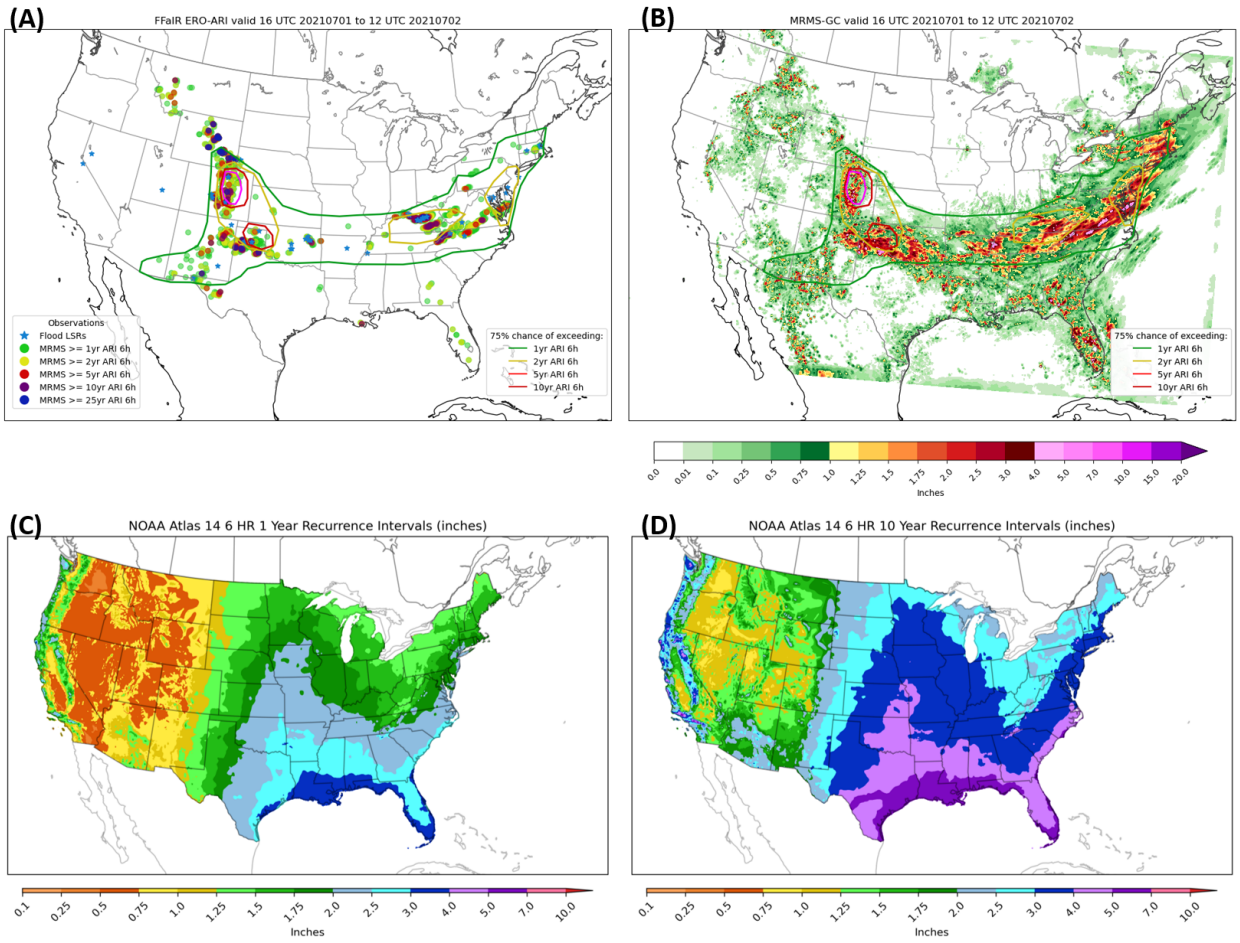


Figure 11: *Example of a subjective verification image for the FFaIR Day 1 ARI-ERO. (B) 20 h MRMS and (A) FFaIR Day 1 ARI-ERO overlaid with flood and flash flood LSRs (blue stars) and the 6 h QPE exceedances of 6 h ARIs for: 1 (green), 2 (yellow), 5 (red), 10 (dark purple), and 25 (dark blue) years, valid 16 UTC 01 July to 12 UTC 02 July 2021. The NOAA Atlas 14 6 h (C) 1 y and (D) 10 y ARI.*

For MRTP verification, when possible the participants evaluated their personal MRTP, when it wasn't they were randomly assigned a MRTP to evaluate. An example of what the verification images for the MRTPs looked like can be seen in Fig. 12A-D. Verification included overlaying the MRTP onto the 6h MRMS QPE, plotting the location of the maximum for rainfall, 6 h ARI exceedance, rainrate[7], and duration of rainfall equal to or exceeding 1 inch per hour. This information was used to evaluate various aspects of the MRTP such as area and orientation; Figs. 13 shows all the aspects participants were asked in the survey to evaluate. The images also included the Critical Success Index (CSI) of the MRTP and the assigned model, as well as the MRTP's Probability of Detection (POD) and False Alarm Rate (FAR). Participants completed the same evaluation on the performance of their assigned model/ensemble 1 inch QPF. The verification images for the models and ensembles were similar to the MRTP images but only the 1in QPF was contoured; see Figs. 12E-F.

[7] Uses MRMS 2min precipitation rate output.

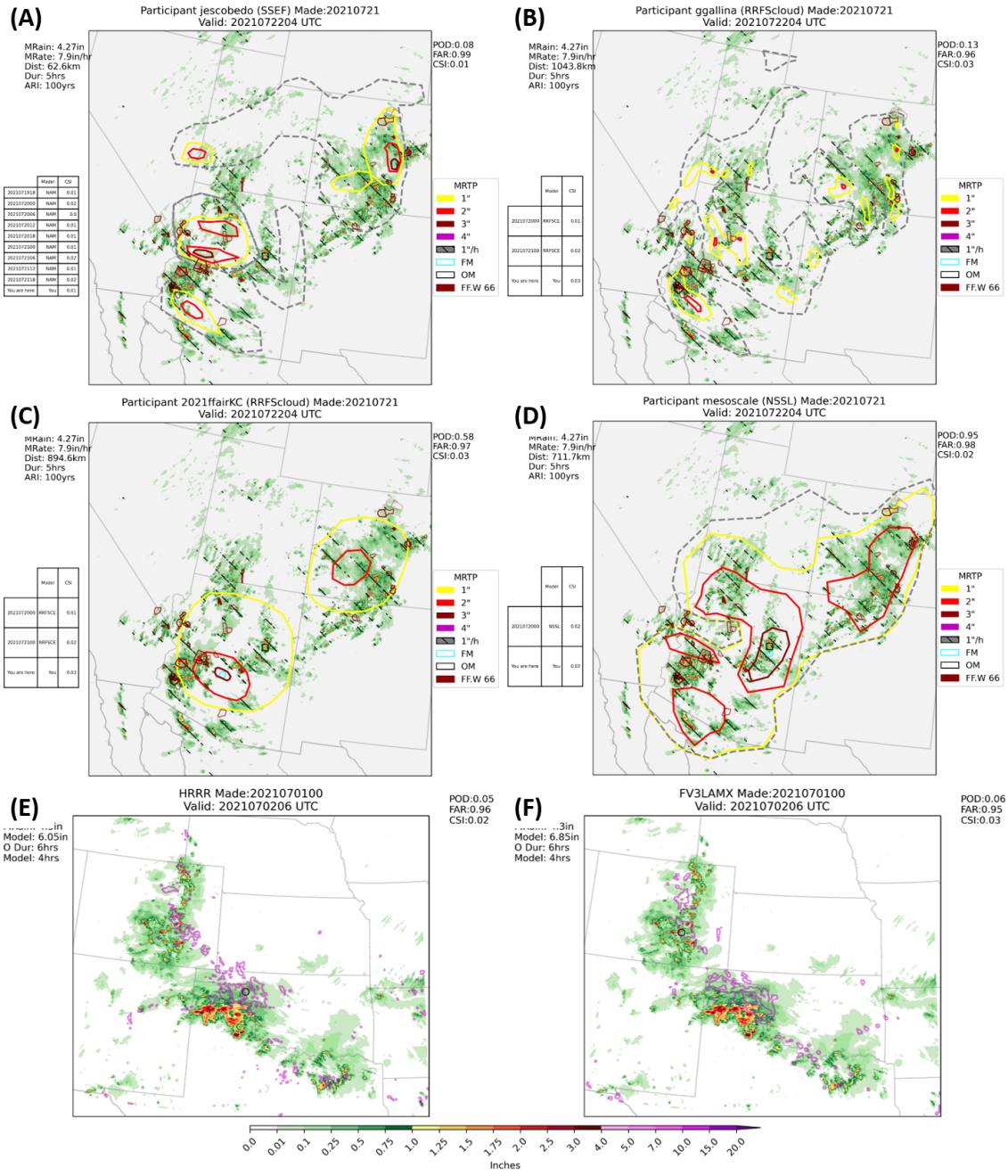


Figure 12: *Example of verification for the MRTPs shown in Fig. 3, and verification for the (E) HRRR and (F) LAMDAX, valid from 00 UTC to 06 UTC July 02, 2021. Shaded is the MRMS 6h QPE. Contours in (A)-(D) are the MRTP forecasts where: yellow - 1 in., red - 2 in., dark red - 3 in., purple - 4 in., and dashed gray - 1in/h rainfall rates. Pink contours in (E)-(F) are the model 1in. QPF. The blue circle is the forecasted location of maximum rainfall. Black objects are the observed location with values listed in the upper left corner for: maximum rainfall (circle), 6 h ARI exceedance maximum (diamond), maximum in rainfall rate (square) and maximum duration of rainfall (star). Red polygons are Flash Flood Warnings issued during the event. On the left side of the images is a table listing the CSI for the forecaster, as well as for each run cycle that was valid during the event for the model/ensemble the forecaster was assigned. Top right of imagest: the forecaster's CSI, POD and FAR scores.*

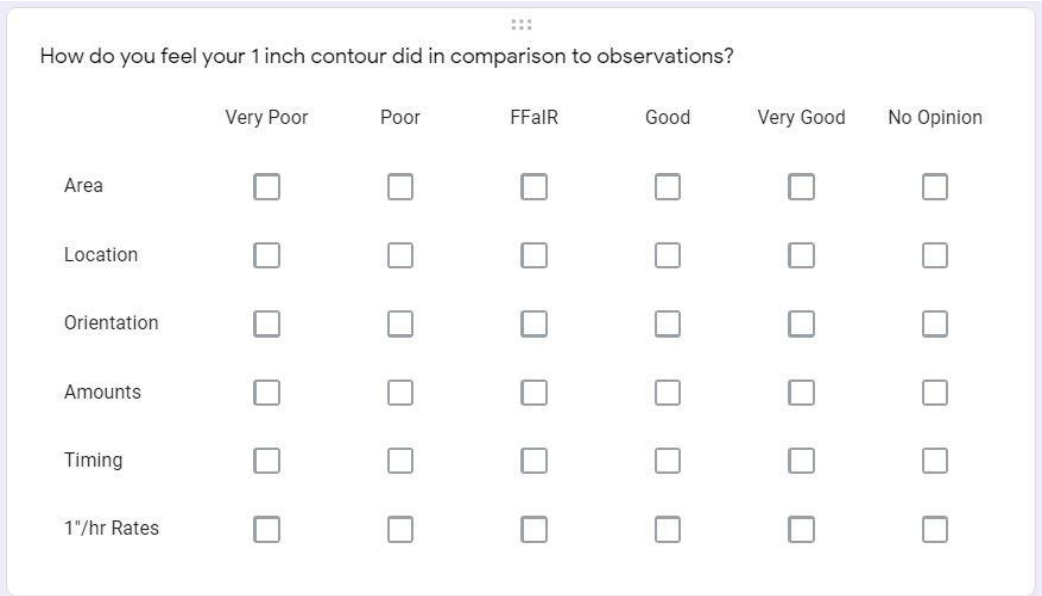


Figure 13: *The aspects of the MRTP that participants were asked to evaluate during verification.*

2.3.2 Objective Verification

As noted in, all QPF verification, including p-rate, were verified against MRMS-GC. MODE, which is part of the Model Evaluation Tools (MET) package[8] (Bullock 2016), was used for daily and experimental long verification. The configuration for MODE is the same configuration that was used last year in FFaIR; refer to Appendix D in the 2020 FFaIR Final Report. As stated in Section 2.3.1, participants were able to use the daily MODE analysis for the 24 h QPF during the verification sessions to help them assess model performance. MODE allows for the visualization of object verification, see Fig 14, along with bulk statistics. During FFaIR, the 24h QPF thresholds assessed were 0.5", 1", 2", 3" and 6". Additionally, both the MRMS and model data are remapped onto a 5km CONUS mask by MODE.

The MRMS's native grid is 1km while the CAMs' grids are roughly 3km. To account for this, the MRMS images made by the FFaIR team remapped the MRMS data to the HRRR grid via a dilated maximum method. This allowed for the maximum in the dataset, which we are often most interested in, to be preserved when losing resolution. This method was used for all the verification metrics and images shown, aside from the ones that used MODE. This also included the 2 min instantaneous precipitation data.

ERO guidance was verified using practically perfect analysis based on the UVFS, which was developed and tested in a previous FFaIR experiment to help with ERO verification (Erickson et al. 2019). The UVFS is a dataset that consists of Local Storm Reports (LSR), STAGE IV exceedance of the 5 yr ARI, STAGE IV exceedance of FFG, and USGS gauge data.

[8] Information on MET can be found at the Developmental Testbed Center website: https://dtcenter.org/community-code/model-evaluation-tools-met.

These are point observations and therefore a 40 km radius of influence is applied to the observations along with a 105 km Gaussian filter (Erickson et al 2021) to create the practically perfect verification. Since the CSU EROs are valid from 12 UTC to 12 UTC and the FFaIR/WPC EROs are valid 16 UTC to 12 UTC, it is possible that one some days the practically perfect analysis used for verification differs slightly from one another.

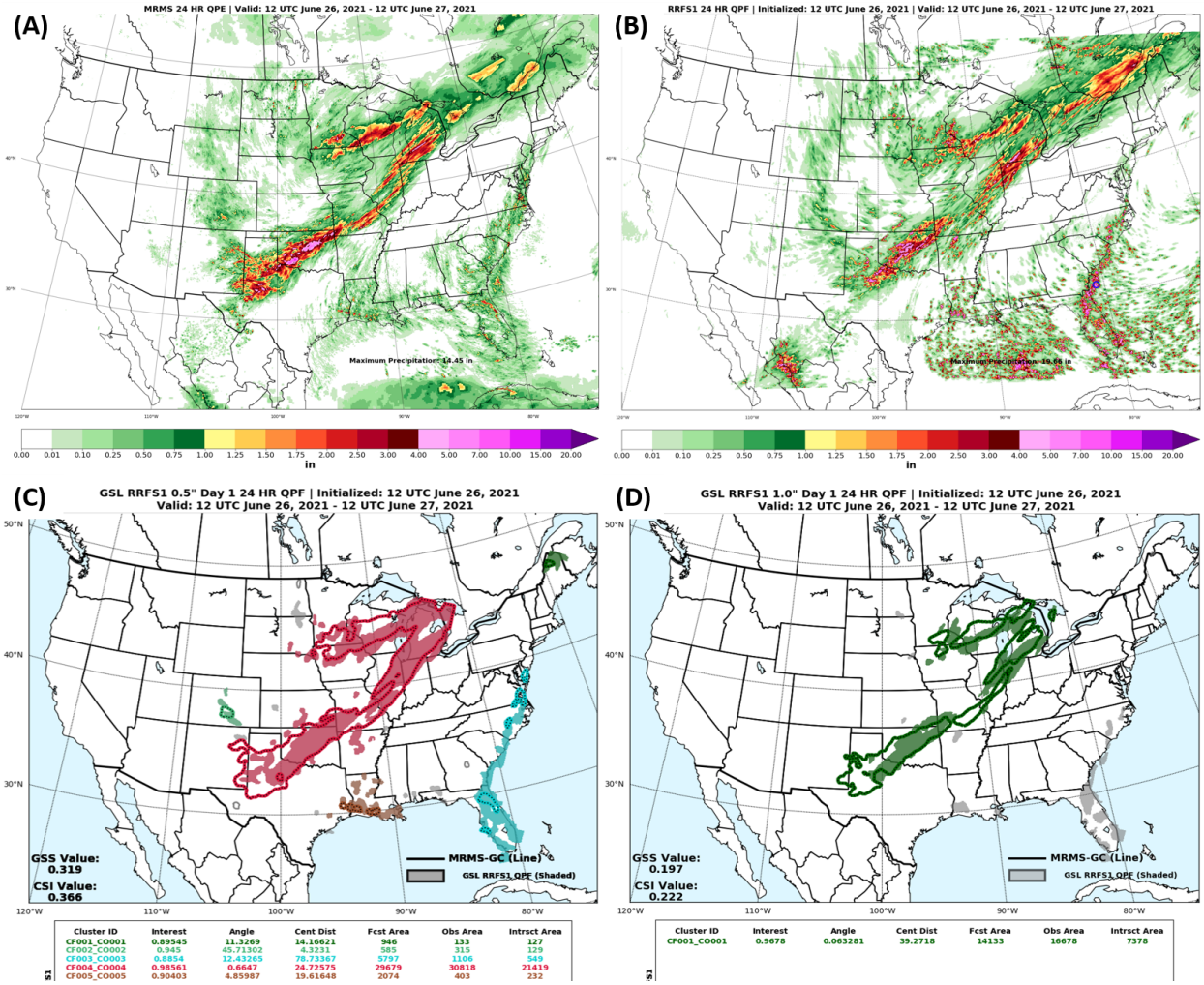


Figure 14: *24 h (A) MRMS QPE and (B) RRFS1 QPF valid 12 UTC 26 June to 12 UTC June 27, 2021. MODE images for RRSF1 at (C) 0.5" and (D) 1" thresholds. The MRMS QPE is contoured and the model QPF is shaded. The differing colors represent different objects identified by MODE. The grey shaded objects indicate objects that were forecasted by the model but had no matching object in MRMS. Below the images is the object id table, it includes information such as forecast area and observed area.*

3. Meteorological Highlights During the Experiment

As has been the trend over the last few years, the 2021 FFaIR Experiment was not lacking in heavy rainfall and flooding events. This section will discuss some of the more impactful events during the experiment. This year, the events in FFaIR were not dominated by

Mesoscale Convective Systems (MCSs), as they were last year, but rather frontal passages and the return of the Southwestern Monsoon. The monsoon itself presented unique challenges and was often the focus of the MRTP exercise. Many of the forecasters commented that they have a newfound respect for the difficulties associated with forecasting during the Monsoon, noting they were glad they did not work in Weather Forecast Offices (WFOs) that were in charge of issuing forecasts for the region.

The first week of FFaIR, June 21-25, saw a deep trough move across the CONUS east of the Rockies, followed by a shortwave moving across the Midwest. The progression of these two features can be seen in Fig. 15, while the EROs for the week can be found in Fig. 16. The most challenging event of the week was a heavy rainfall event across northern Missouri that occurred on June 25th, that resulted in widespread flooding and numerous water rescues along and north of Interstate 70. A system from the previous night was forecast to leave a boundary draped across the Nebraska/Missouri region with a low moving into the region from Arizona. The interaction between the two presented a difficult forecast, so much so that the participants were tasked with making a Day 2 and Day 1 MRTP for the event, both valid at 09 UTC 25 June 2021.

Figure 17 shows three participant's Day 2 (left) and Day 1 (right) MRTPs for the event. Unfortunately, data from all three EMC FV3-LAMs was missing, therefore the forecasts relied on the operational CAMs and the experimental deterministic members from the SSEF. At Day 2 the models were in two different camps, with the SSEF members' axis of heavy rainfall further north into Nebraska, while the operational models had the axis further south from eastern Kansas into Missouri. The same was also true of the HREF and the RRFSCE, where the experimental ensemble had a northern axis. These differences between the guidance can be seen in Fig. 18.

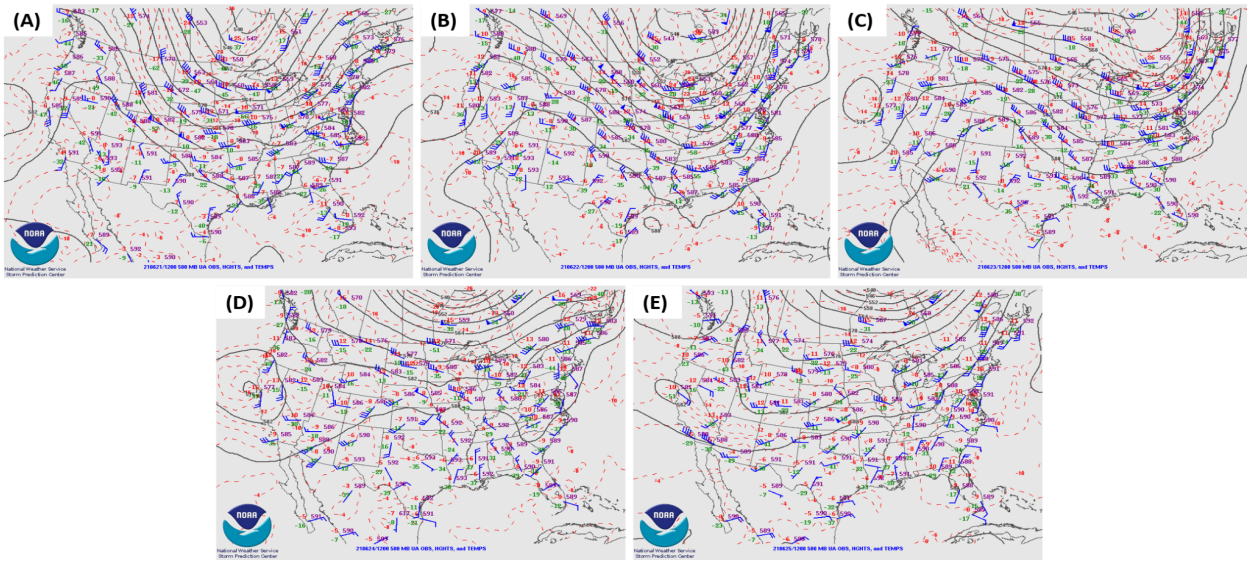


Figure 15: *SPC 12z 500mb analysis for (A) June 21, (B) June 22, (C) June 23, (D) June 24 and (E) June 25, 2021.*

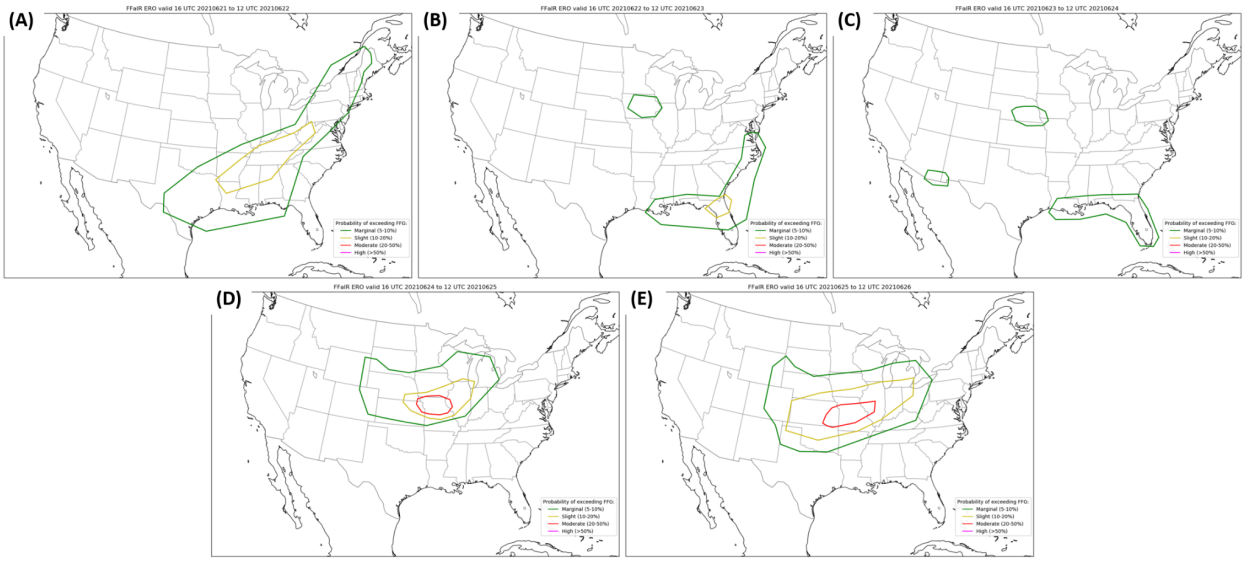


Figure 16: *Experimental EROs issued by the Week 1 participants, valid: (A) 16 UTC 21 June to 12 UTC 22 June, 2021, (B) 16 UTC 22 June to 12 UTC 23 June, 2021, (C) 16 UTC 23 June to 12 UTC 24 June, 2021, (D) 16 UTC 24 June to 12 UTC 24 June, 2021, and (E) 16 UTC 25 June to 12 UTC 26 June, 2021. ERO Risks: Marginal - green, slight - yellow, moderate - red, and high - pink.*

Although at first it seemed to identify a shortcoming of the experimental models and ensembles at longer lead times, that assumption is not entirely accurate. The SSEF members, NAMnest, and the RRFSCE are run out to 60+ hours while the HRRR and HREF are only run out to 48 hours. Additionally the experimental guidance available is only run at 00z and do not have data assimilation. On the other hand, the 06z and 12z forecasts from the operational models were available, with the 12z forecast needing to be used to cover the valid forecast time for the HRRR and HREF at Day 2[9]. This meant that the experimental guidance was initialized on the 18z GFS run from the previous day, which also had a northern bias (not shown). This likely was a large contributor to the incorrect location of the axis of heavy rainfall in the experimental guidance at Day 2. If the EMC FV3-LAMs, which are run at 00z and 12z out 60 hours, had not been missing, a comparison between 00z and 12z experimental guidance could have been done to determine how great of an impact the GFS initial conditions had on the long range forecast. Especially since the 06z GFS, which is used to initialize the 12z FV3-CAMs, had shifted the axis of rainfall further south.

Figure 19 shows the EROs issued by the FFaIR participants during week 2 of FFaIR, June 28 - July 2. The week revolved around two tropical systems and a front that slowly lagged south across the eastern CONUS during the latter half of the week. Tropical Storm Danny impacted the Carolinas into the Mid-Atlantic while Tropical Storm Enrique moved across the southern Baja Peninsula (Fig. 20A). The latter of the two funneled moisture into UT/AZ on June 28th, helping to drive the extreme flooding that was seen at Zion National Park in UT. The

[9] Forecast hour 57 if you use the 20210623 00z run.

flooding damaged roads across the park, see Figs. 20B-C and resulted in the closure of State Route 9 into the park.

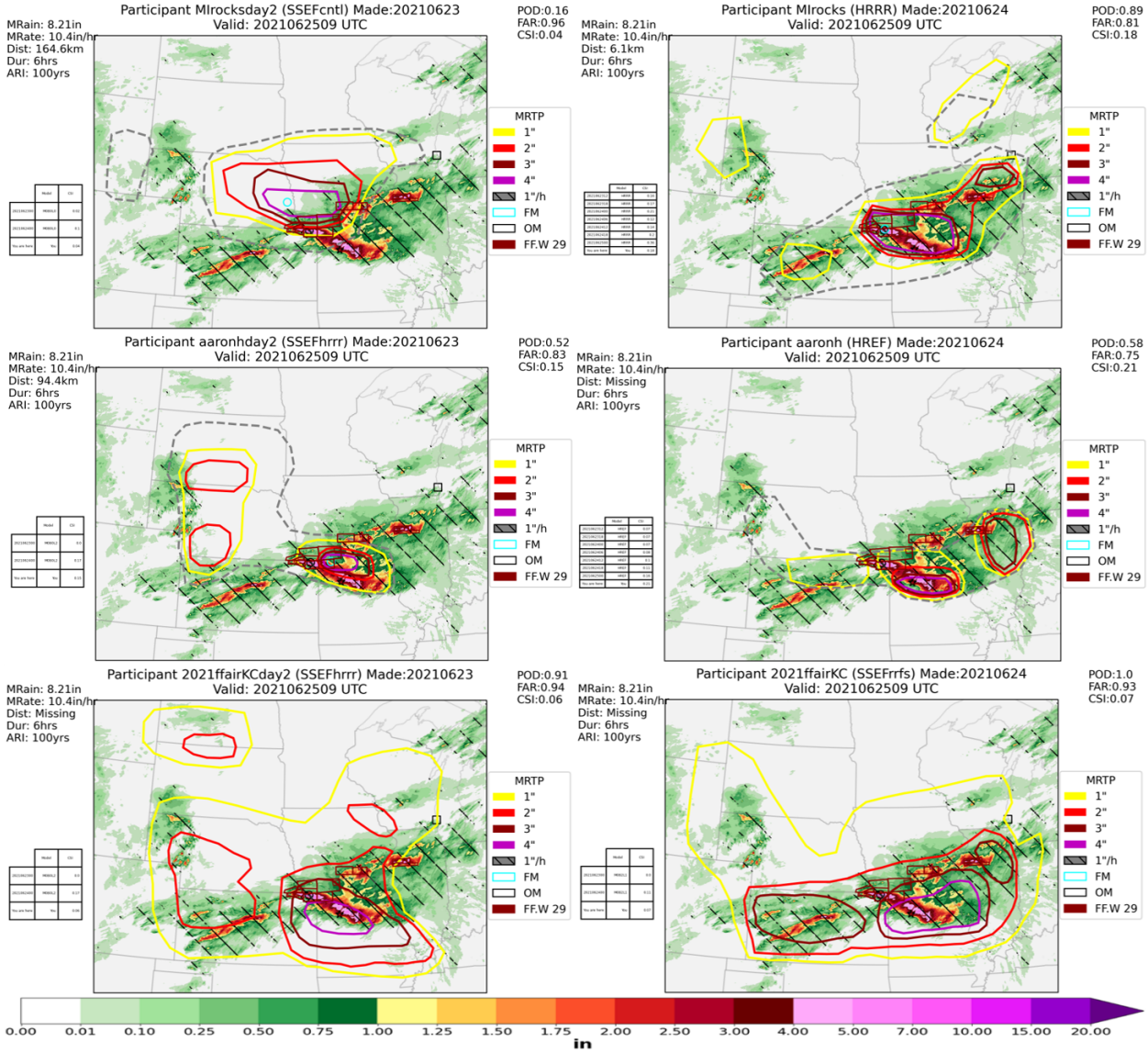


Figure 17: *Day 2 (left) and Day 1 (right) MRTPs valid 03 UTC to 09 UTC 25 June 2021, issued by MIrocks (top), aaronh (middle), and ffairKC (bottom). Refer to Fig. 12 for the description of the MRTP graphic.*

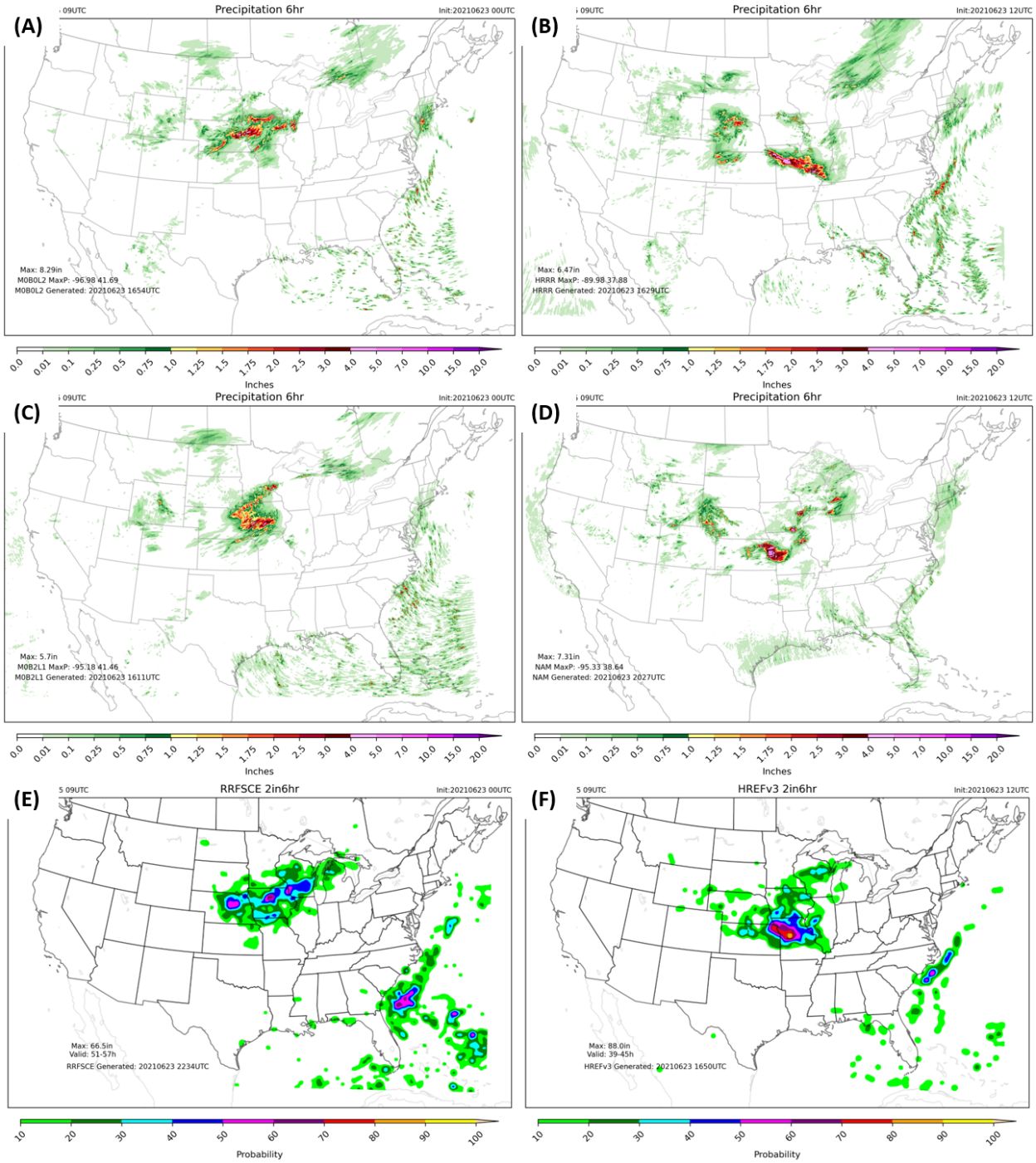


Figure 18: *6 h QPF from the 00 z 20210623 cycle of the (A) HRRR-like member, (B) HRRR, (C) RRFS-like member, and (D) NAMnest and the probability of exceeding 2 inches in 6 h from the (E) RRFSCE and (F) HREF valid for the same time as Fig. 17.*

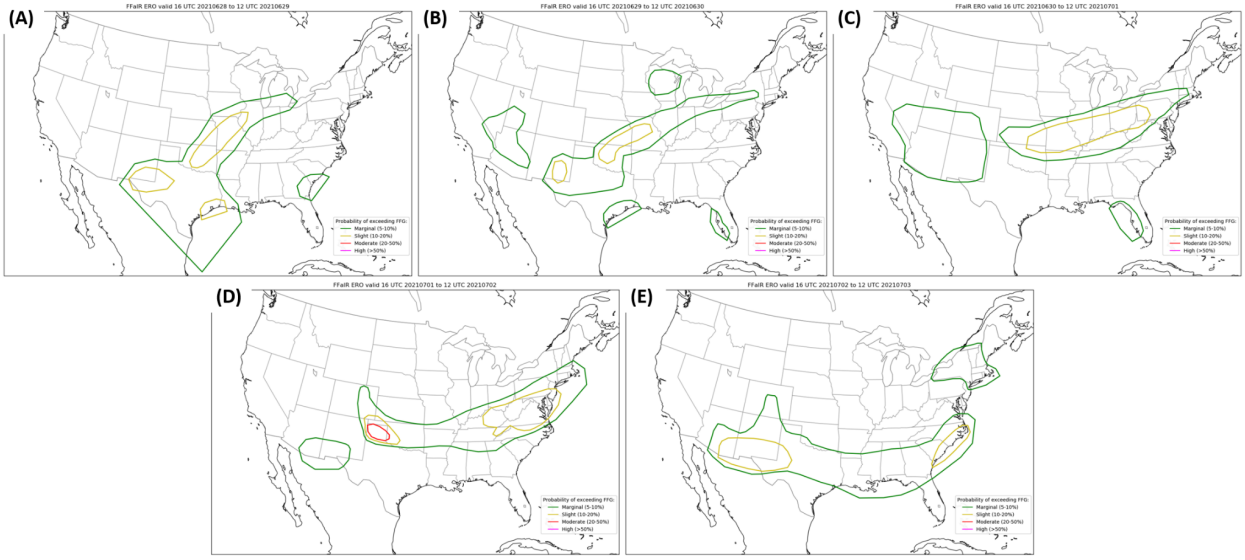


Figure 19: *Same as Fig. 16 but for Week 2. (A) 16 UTC 28 June to 12 UTC 29 June, 2021, (B) 16 UTC 29 June to 12 UTC 30 June, 2021, (C) 16 UTC 30 June to 12 UTC 01 July, 2021, (D) 16 UTC 01 July to 12 UTC 02 July, 2021, and (E) 16 UTC 02 July to 12 UTC 03 July, 2021.*

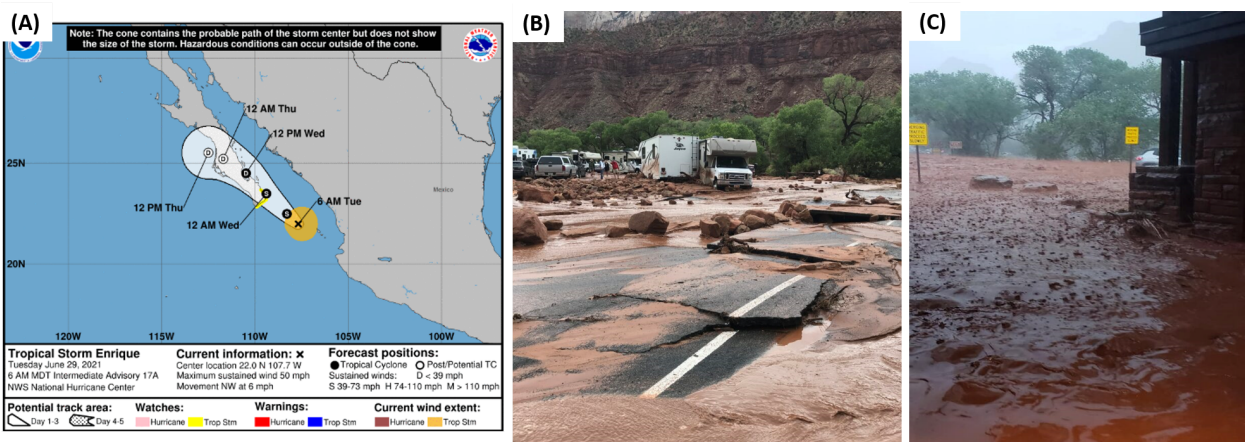


Figure 20: *(A) The National Hurricane Center's forecast for Tropical Storm Enrique value June 29, 2021. (B) Picture of the damage from the flood in Zion Park and (C) picture of the flooding at the gate of Zion Park on June 28, 2021 (courtesy of article written by Reed, C. (2021))*

On the afternoon of July 1, the slow-moving front became nearly stationary until the following morning; see Fig. 21. This resulted in widespread flooding from the Mid-Atlantic to Texas Panhandle and across the front range of the Rockies. The FFaIR ERO verification valid 16 UTC July 1 to 12 UTC July 2, 2021 can be seen in Fig. 10. The practically perfect (Fig. 10C.) suggests a Moderate Risk of excessive rainfall leading to flooding was warranted along much of the front. In Greeley, CO (located between Boulder and Fort Collins) approximately four inches of rain fell in an hour, resulting in widespread flooding across the city[10]; see Fig. 22. In Oklahoma City, storms trained over the city, resulting in WPC issuing an MPD for the area and

[10] https://denver.cbslocal.com/2021/07/01/street-flooding-greeley-heavy-rain-standing-water/

numerous videos of flooding across the metro were shared via Twitter from the many meteorologists in the region; Fig. 23.

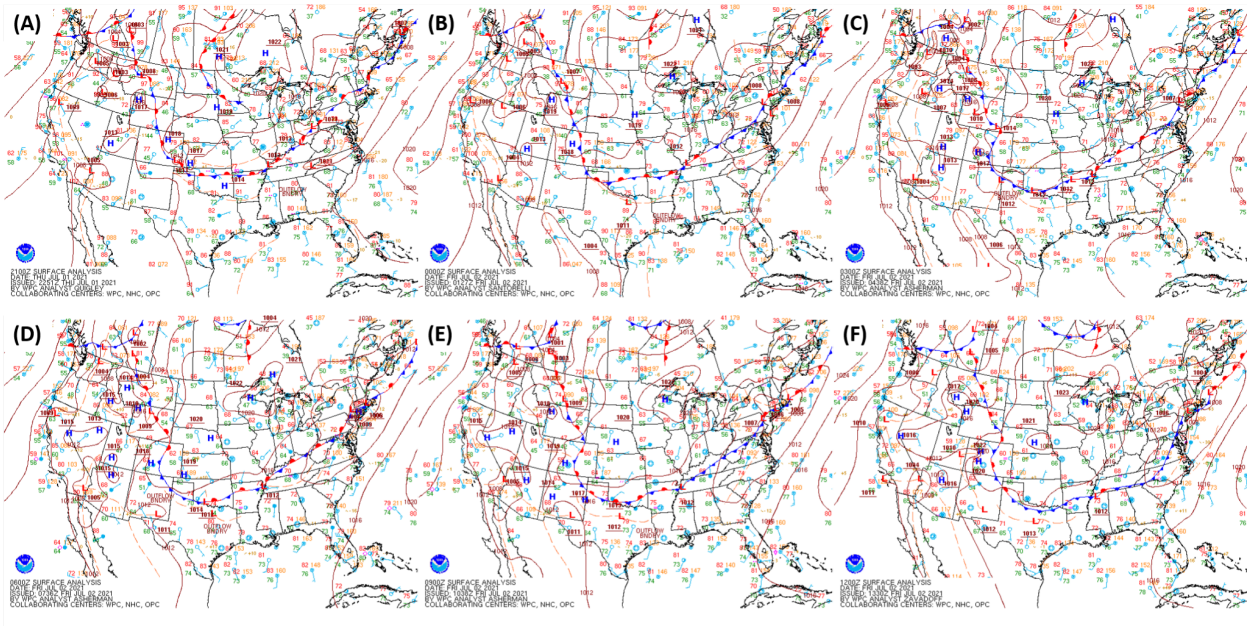


Figure 21: *WPC's surface analysis valid (A) 21 UTC 01 July 2021, (B) 00 UTC, (C) 03 UTC, (D) 06 UTC, (E) 09 UTC, and (F) 12 UTC 02 July 2021.*

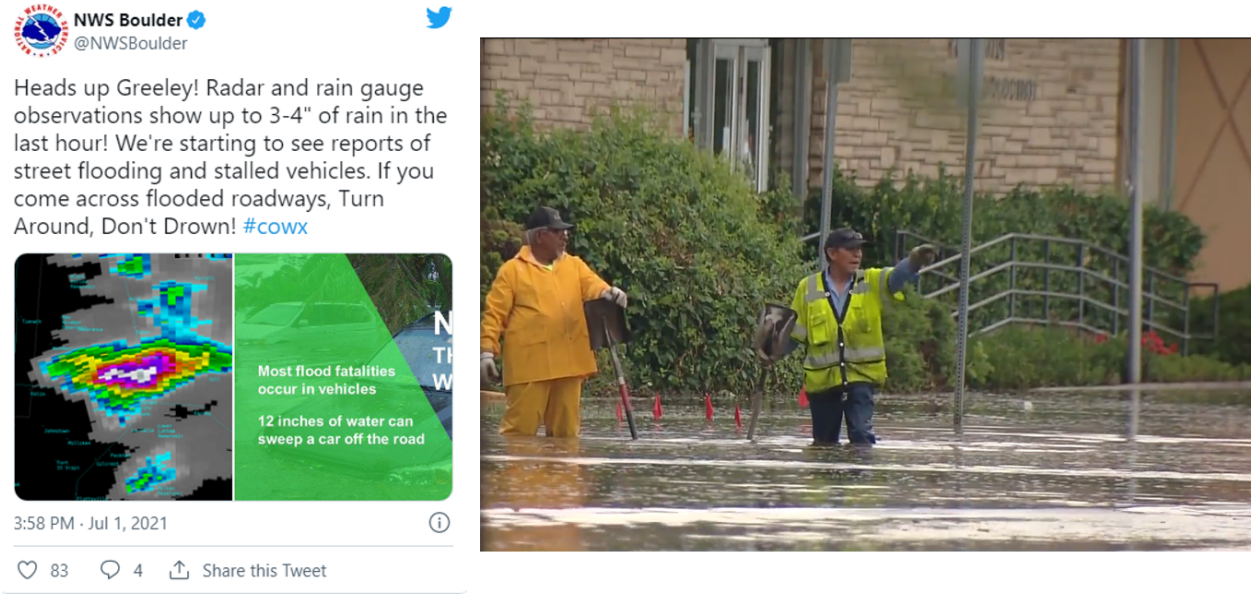


Figure 22: *(A) Tweet from the Boulder NWS about the heavy rainfall occurring in the town of Greeley, CO on 01 July 2021. (B) Image of the flooding in Greeley, courtesy of article written by Sarles, J. (2021).*

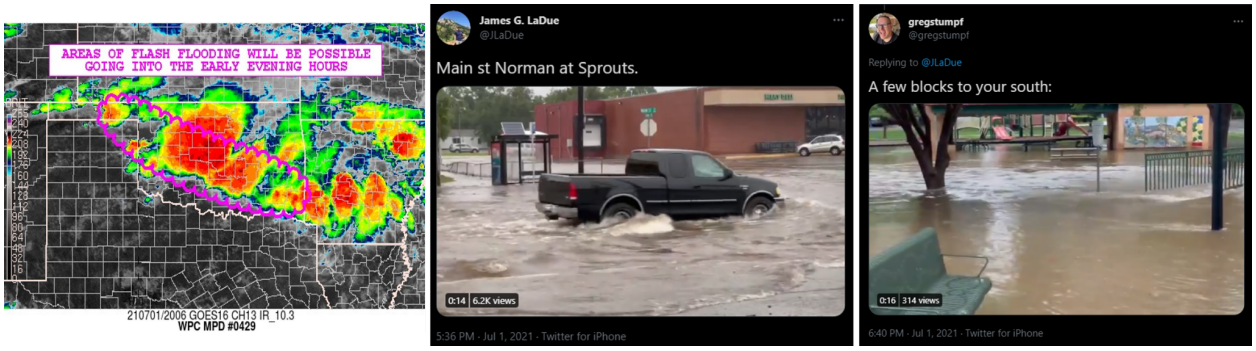


Figure 23: *Left: The MPD issued by WPC for flooding potential in Oklahoma City valid at 2006 UTC 01 July 2021. Right two images are pictures taken by Storm Prediction Center employees of the flooding.*

The second half of FFaIR was dominated by the Southwestern Monsoon. As can be seen in Figs. 24 and 25, there was not one day during this time that the FFaIR participants did not issue at least a Slight Risk somewhere in the southwest for their ERO. Although the monsoon was at the forefront of the discussion and will be discussed in further depth below, it was not the only impactful heavy rainfall/flooding event that occurred. On July 12, along the southern border of Pennsylvania and New Jersey a flash flood emergency was issued and the WPC MPD issued for the event included a considerable flooding tag; see Fig. 26A-B. Portions of Bucks and Burlington Counties received 6 to 10 inches of rainfall in 3 to 4 hours[11]. This resulted in widespread flooding, water rescues, and people having to be evacuated from their homes (see the LSRs in Fig. 26C).

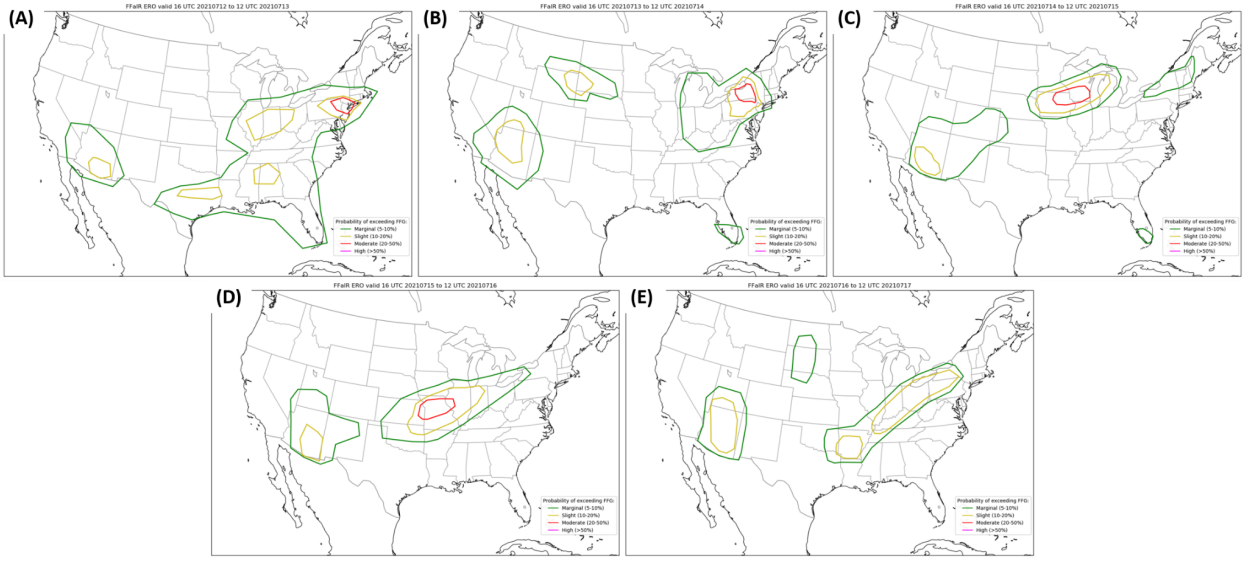


Figure 24: *Same as Fig. 16 but for Week 3. (A) 16 UTC 13 July to 12 UTC 14 July, 2021, (B) 16 UTC 14 July to 12 UTC 15 July, 2021, (C) 16 UTC 15 July to 12 UTC 16 July, 2021, (D) 16 UTC 16 July to 12 UTC 17 July, 2021, and (E) 16 UTC 17 July to 12 UTC 18 July, 2021.*

[11] https://6abc.com/flash-flood-emergency-bucks-county-flooding-nj-bristol-pa/10882406/

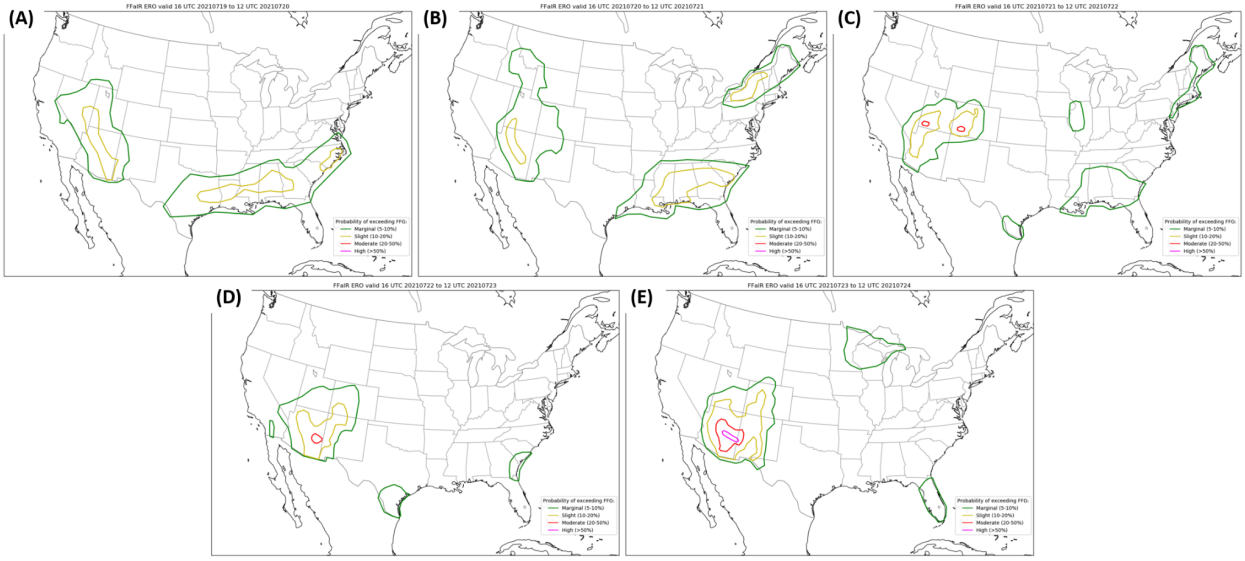


Figure 25: *Same as Fig. 16 but for Week 4. (A) 16 UTC 19 July to 12 UTC 20 July, 2021, (B) 16 UTC 20 July to 12 UTC 21 July, 2021, (C) 16 UTC 21 July to 12 UTC 22 July, 2021, (D) 16 UTC 22 July to 12 UTC 23 July, 2021, and (E) 16 UTC 23 July to 12 UTC 24 July, 2021.*

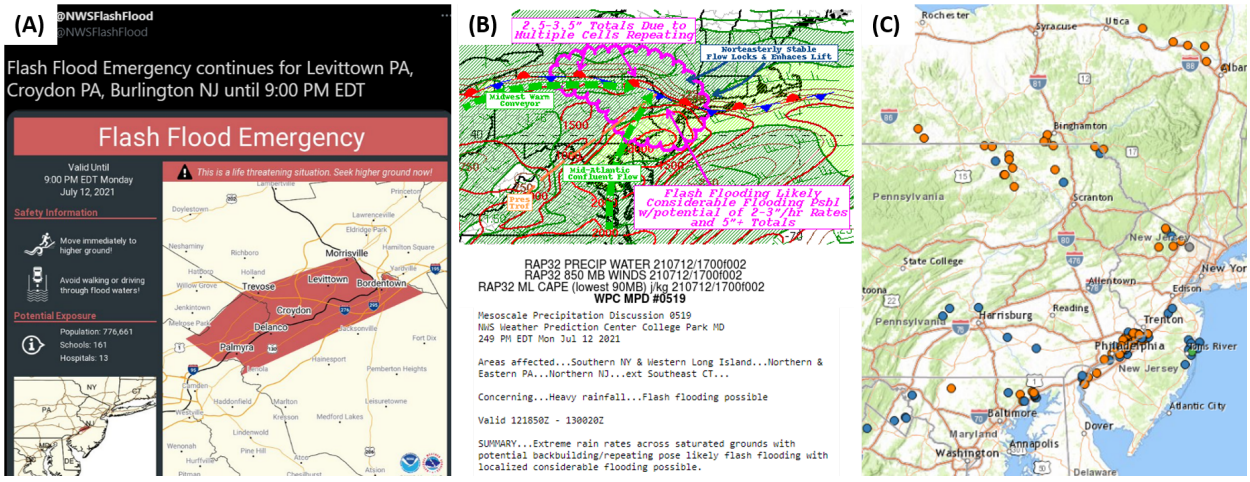


Figure 26: *(A)The alert for the Flash Flood Emergency issued along the southern border of PA and NJ on 12 July 2021. (B) The MPD issued by WPC highlighting the likelihood of flash flooding for the event. (C) Local Storms Reports of heavy rainfall and flooding.*

The 24 h QPF for the PA/NJ even from various models, valid from 12 UTC July 12 to 12 UTC July 13, can be seen in Fig. 27. Overall the models, both operational and experimental, hinted at heavy rainfall across this area. The RRFS1 (Fig. 27E) even had the location of the heavy rainfall correct. That said, many of the participants noted that the RRFS1 had numerous areas with 4+ inches of rain being forecasted, so it would be impossible to know which area to

focus on as the largest threat. For instance, one participant wrote "RRFS actually captured the Philadelphia heavy rain, although it also had similar spurious floods elsewhere in the NE." Another noted, "The places where very heavy rain did occur in NY/PA and TX were generally pretty good, it was the over-forecasting in these other areas that was problematic." These comments show that just because a model gets one aspect of a forecast correct, it does not mean the forecast was necessarily useful.

On July 14th, the first of two flooding events across the Albany, NY region in less than a week occurred. Over the course of a couple of hours, towns just east of Albany saw 2 to 4 inches of rain, with the largest total in Averill Park with 4.5 inches. Figure 28A, a tweet sent out from NWS Albany, shows how isolated the heavy rainfall was. The flooding washed out numerous roads across the area, caused a landslide on Route 66 and resulted in schools having to close to clean up the mess left behind[12]; see Figs. 28B-C. Then on July 19th, a nearly stationary storm developed near the town of Fonda, which is located west of Albany. It remained stationary for nearly two hours, resulting in 2 to 4 inches of rainfall over an isolated area; see Fig. 29A. Fonda, which sits along the Mohawk River with higher terrain from the Adirondacks surrounding the town, is used to flooding, but this event was not typical according to the residents of the town. The unusualness of the flood was likely aided by the mud that came down from the bluffs as the creek above overflowed its banks, resulting in storm drains being clogged[13]. A friend of the FFaIR team happened to be driving into Fonda on his way to work that day and sent this the following day when asked about the flood: " (I) saw all the water escaping from the hillsides as the water table surged. I stopped and turned around. Literally every other car I was driving with drove into Fonda anyway. I drove through the town just this morning. All the cars I was on the road with I could see on the side of the road....totaled." The event resulted in a State of Emergency being declared, exits on the New York Thruway being closed, and the main street running through town "turning into a river"; see Figs. 29B-C.

[12] https://www.timesunion.com/news/article/Heavy-rains-flooding-close-roads-in-Rensselaer-16315342.php
[13] https://dailygazette.com/2021/07/20/fonda-hit-by-flash-flooding-monday-night-village-cleaning-up/

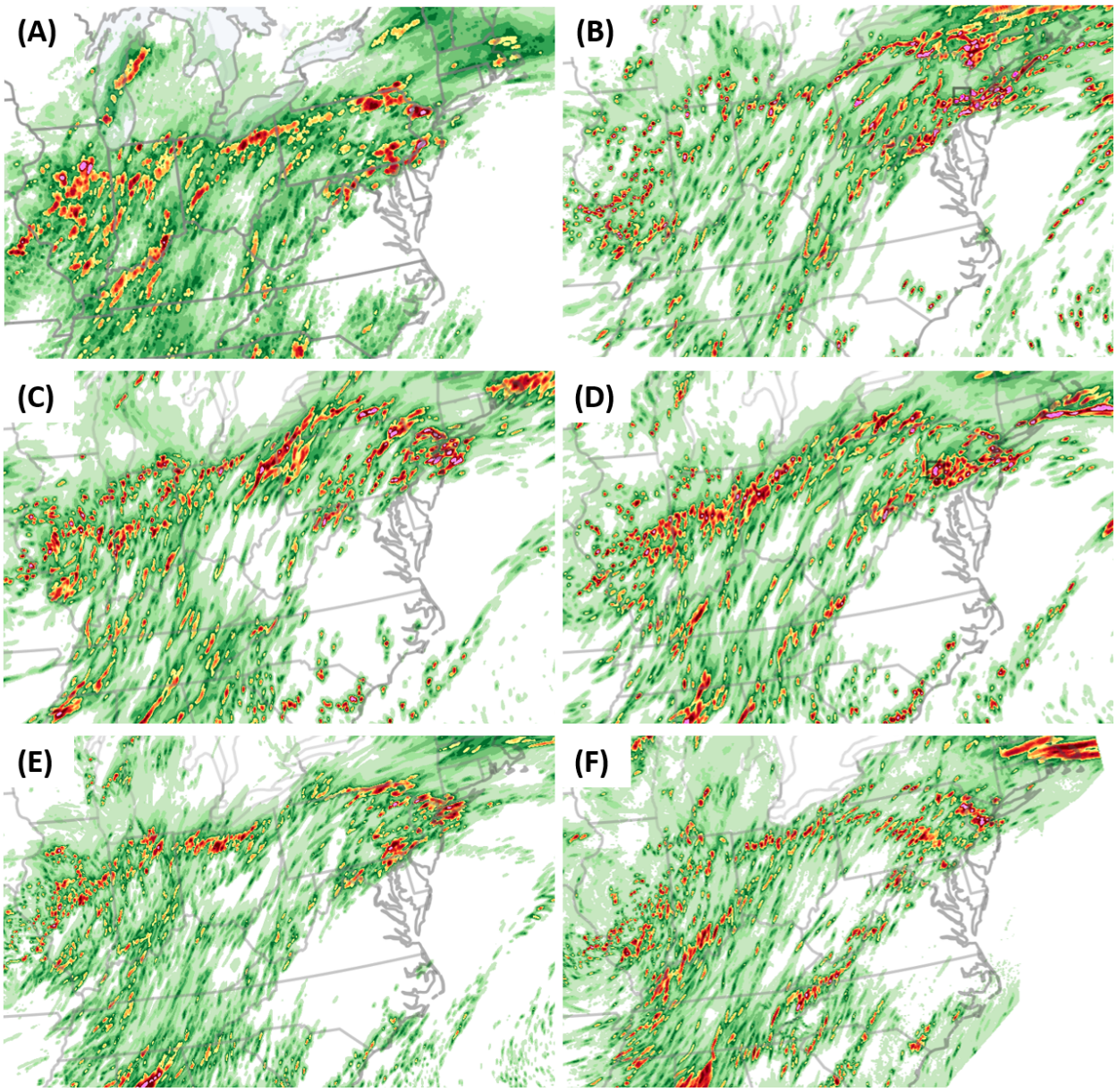


Figure 27: *24 h (A) MRMS QPE and (B) RRFS1, (C) LAMX, (D) LAMDAX, (E) HRRR, and (F) NAMnest QPF valid 12 UTC July 12 to 12 UTC July 13. If an image has a blue circle or grey square it is indicating that is the location of the model's maximum QPF.*

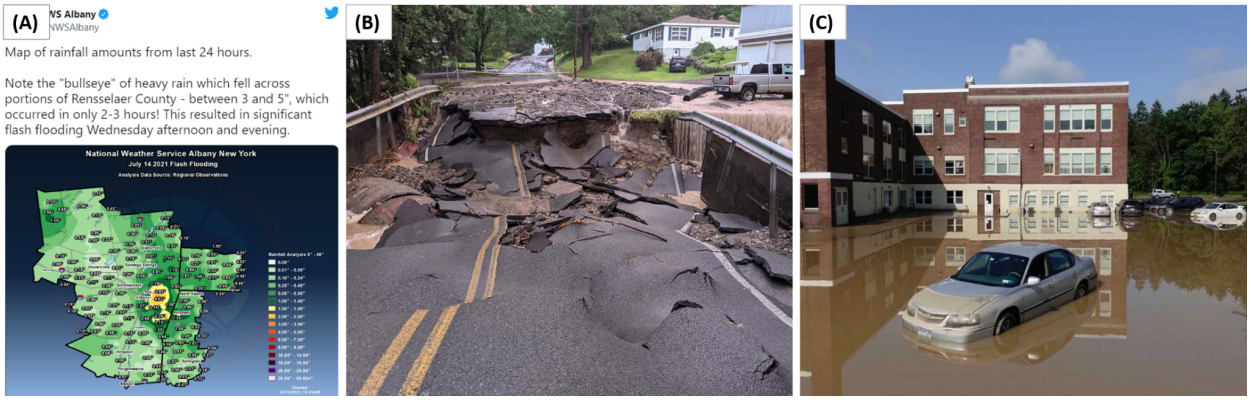


Figure 28: *(A) Tweet from the WFO Albany highlighting the isolated nature of the heavy rainfall that occurred on July 14, 2021. (B)-(C) Images of the damage from the flash flooding that occurred as a result of the heavy rainfall east of Albany, NY; courtesy of the article written by Crowe and DeMola (2021).*

Figure 29: *(A) 6 h MRMS valid at 00 UTC 20 July 2021. (B)-(C) Images of the flooding in Fonda, NY; courtesy of the articles written by LaPointe, S. (2021) and WNYT Staff (2021).*

3.1 Southwestern Monsoon

After the southwestern monsoon being relatively weak the past two years, the monsoon this year was in full force. Figure 30 shows the accumulated rainfall across the southwestern part of the CONUS during weeks 3 and 4 of FFaIR (July 12-24) compared to the accumulated rainfall during weeks 3 and 4 of FFaIR 2020 (July 6 - 18) and the same date span of dates for this year's week 4 but valid in 2020. As can be seen in Fig. 30A, a broad area of 4+ inches of rain fell across AZ as well as northwestern NM and southwestern UT into southern NV, with parts of central AZ seeing over 10 inches, during these two weeks. To say that the monsoon kept the participants on their toes and challenged their forecasting skills would an understatement. During week 4 of FFaIR, 3 of the 5 MRTP domains were over the four corners region. By the end of the week, the week 4 participants were commenting on how difficult it was to forecast for these events and how they have a new found respect for the forecasters at the WFOs in the region.

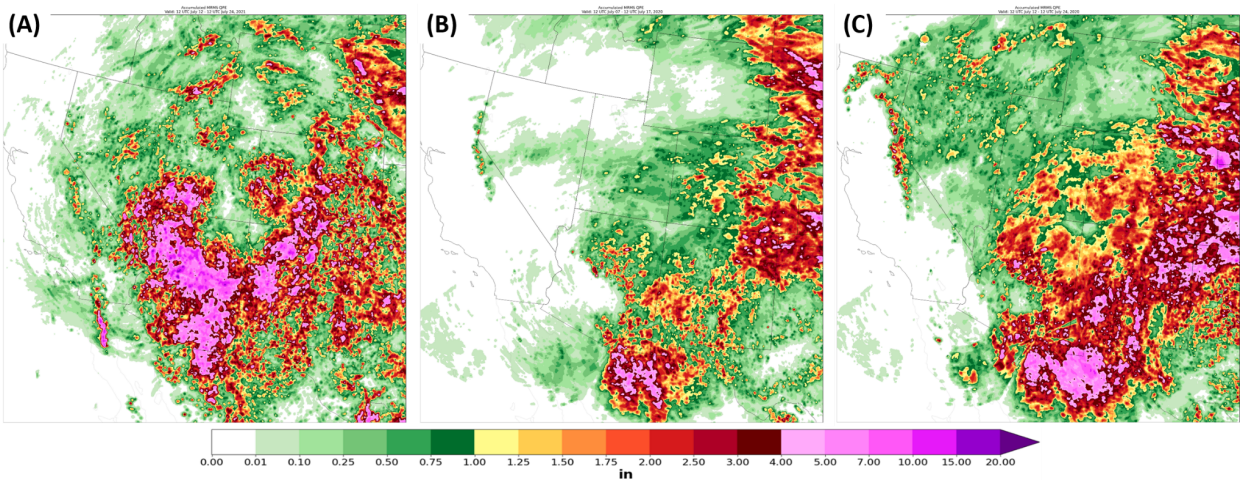


Figure 30: *Accumulated MRMS totals for (A) 12 July to 24 July, 2021, (B) 06 July to 18 July, 2020, and (C) 12 July to 24 July, 2020.*

Another topic of discussion was about the different approaches taken in drawing the MRTPs for the isolated events, specifically if highlighting a large area of 1 to 2 inches so you capture the isolated events was best (i.e. having high POD but also high FAR) or if trying to draw many small areas (i.e. low FAR but also low POD) was better. Figure 31 shows the variety of approaches taken by participants when forecasting for the monsoon. As can be seen, some forecasters used the 1 inch contour more as a way to encompass any rain. For instance, one forecaster commented: "Also considered terrain features such as Wasatch Mountains in Utah, San Juans in Colo, and Arizona plateau near Flagstaff. Tried to hit an envelope of 1-inch, going for high POD at the expense of FAR." Others used the 1in/h contour to try and capture anywhere there was a chance of isolated 1+ inches of rain and then used the precipitation contours to try and hone in on areas where the highest precipitation seemed likely. In some cases (i.e. Fig. 31B) the forecaster tried to be extremely precise. Still others changed their methodology for capturing the events throughout as the week [progressed. By the end of the week, the participants were exhausted from difficulty of forecasting in the region, with comments like these by Friday:

> "Trying to follow up a good day yesterday, I am sure this won't end well. I think there will be some very isolated high rain amounts after this 6hr period. I tried to target those with 2" areas since it's nearly impossible to nail a 3" or 4". I threw a few 3"s out there. I'm not really confident on the area. It would have been easy just to circle big areas. My brain is dead so I just kinda threw my hands in the air. Tucked a few small polygons in CO since I think there will be a few isolated 2" spots (like 2-4 total) so I tried to hit the surrounding 1-1.99 associated with them."

> "I have a whole new respect for WPC. How do they not get burned out with all the guidance overload? I keep staring at my MRTP, and like "heck if i know?!" "

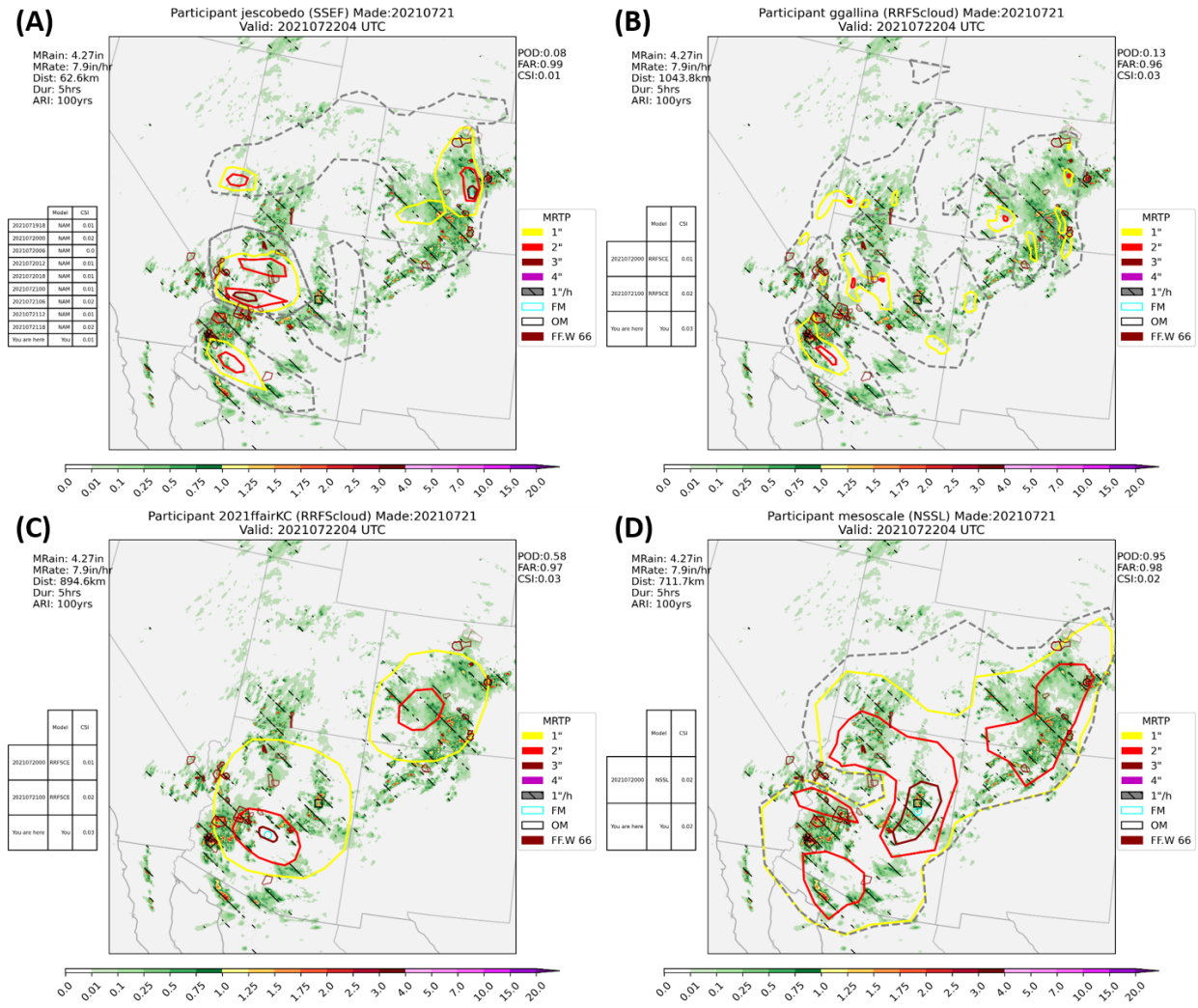


Figure 31: *Refer to Fig. 12 for a description of the MRTP images. MRTP forecasts valid 22 UTC 21 July to 04 UTC 22 July 2021 from (A) jescobedo, (B)ggallina, (C) 2021ffairKC, and (D) mesoscale.*

As can be imagined with the amount of rain seen from July 12th to the 24th, each day there were numerous impacts across the region. On July 15th in UT, heavy rainfall resulted in flooding that led to the derailment of a train (Fig. 32A). The floodwaters surrounding the location of the derailment made it impossible to reach the passengers for several hours[14]. Figure 33 shows the evolution of the event, with convection developing at 21 UTC July 15 and slowly meandering south/southwestward until 0230 UTC July 16 when a progressive line of convection out of NV merged with the cell and finally was pushed out of the area by 04 UTC. This resulted in an isolated area, just to the northwest of Lund, UT seeing over 6 inches of rainfall (Fig. 32B); MRMS maximum rainfall for the whole CONUS was 9.53 inches and was recorded at this location, see the blue circle in Fig. 34A.

14 https://www.fox13now.com/news/local-news/rain-and-rising-floodwaters-play-role-in-southern-utah-train-derailment

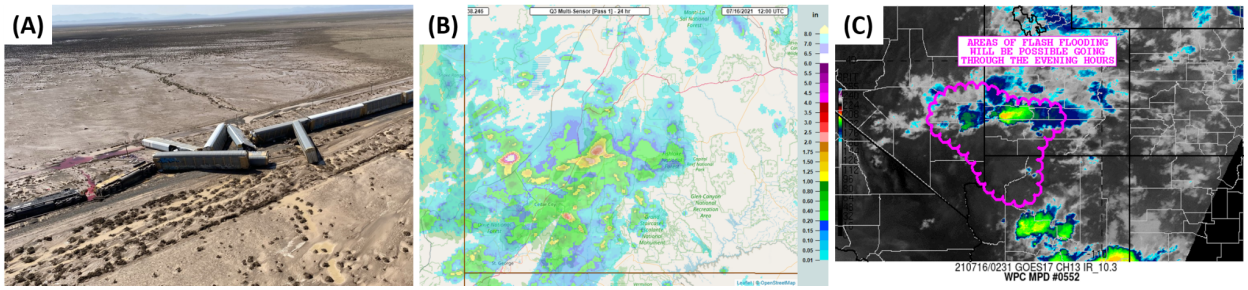


Figure 32: *(A) Image of the train derailment in UT; courtesy of article written by FOX 13 News (2021). (B) MRMS 24 h QPE valid 15 July to 16 July, 2021. (C) MDP #0552 issued by WPC at 0231 UTC 16 July, 2021.*

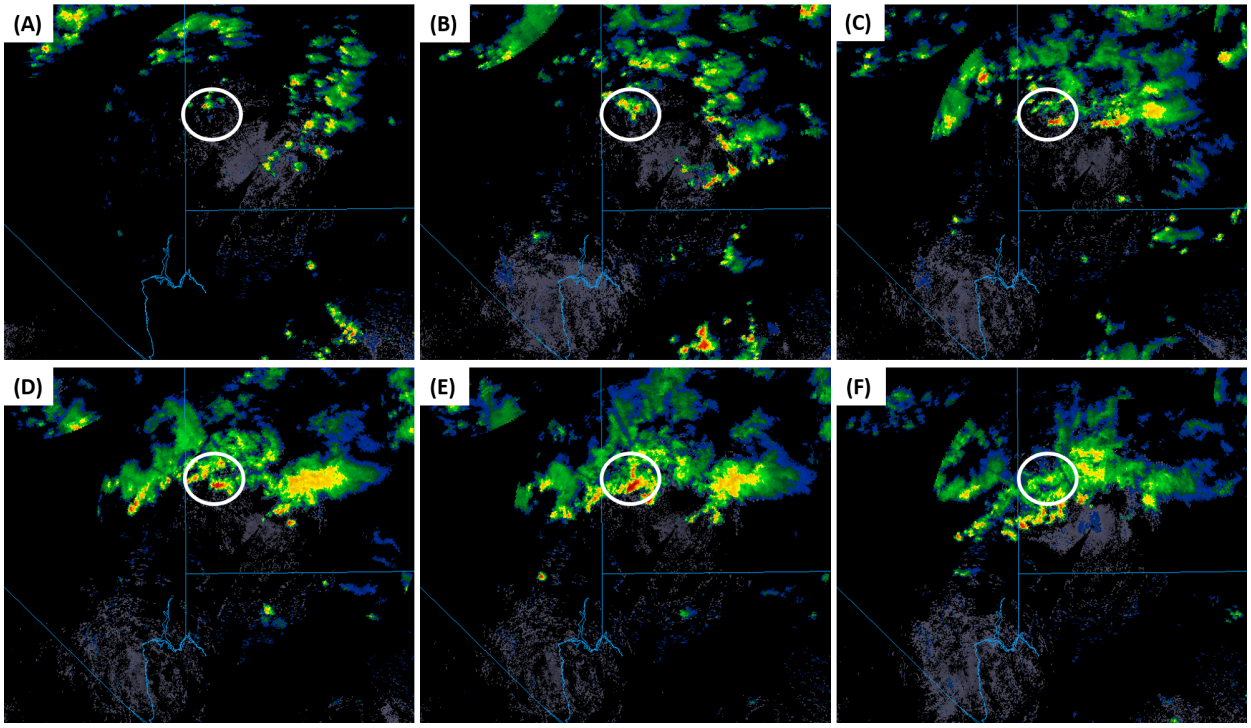


Figure 33: *Composite reflectivity valid: (A) 2100 UTC and (B) 2300 UTC 15 July 2021 and (C) 0100 UTC, (D) 0230 UTC, (E) 0300 UTC, and (F) 0400 UTC 16 July 2021.*

Although both experimental and operational models indicated rainfall across southwestern UT, the potential for isolated areas of heavy rainfall was not consistent among the models. Figure 34 shows a variety of forecast outcomes, with the HRRR, RRFS1, LAMX, and the SSEF deterministic members all forecasting light precipitation across the region; Figs. 34B-E. The LAMDAX and the NAMnest (Figs. 34F-G) both had isolated areas of heavy rainfall exceeding 2 inches in the area of interest but also had similar isolated convection to the south across the Grand Canyon area, which did not verify. Both the HREF and RRFSCE each had low probabilities of 2 inches in 6 hours over the area (not shown) but at the higher threshold of 3 inches in 6 hours (Figs. 34H-I), the RRFSCE had a small 10% contour in the vicinity while the HREF had nothing. The wide plethora of outcomes in the models/ensembles show not only the

difficulty of forecasting such highly localized events but also highlight the necessity of mesoanalysis in the forecasting process and products such as the WPC MPD (like the one issued for this event; see Fig. 32C) to identify these isolated but extreme rainfall events.

The wildfire season in the western U.S. has been very active over the last few years, resulting in an abundant number of burn scars across the region. Even a relatively light amount of rainfall (<0.5") over a burn scar can result in flash flooding and debris flow. This, combined with the active monsoon, led to an elevated risk of flash flooding associated with rainfall, especially since heavy rainfall was not necessarily needed to drive the threat. For instance, Flagstaff, AZ was impacted by flash flooding and debris flows two days in a row, July 13th and 14th. On July 13th, heavy rain fell to the east and north of the city, with the precipitation to the north falling over the Museum Fire burn scar. NWS Flagstaff estimated that 1 to 2 inches of rainfall occurred in 30 to 45 minutes over the scar[15] (Fig. 35A). This created a debris flow that quickly moved downhill into portions of northern Flagstaff (Fig. 35B-C). The following day, rain once again fell over the burn scar, though the heaviest was east of the heaviest rain from the previous day (Fig. 36A). Significant rainfall also occurred over portions of the city, where radar indicated rain rates ranged from 4-8 in/h, with rainfall in some portions of the city seeing 2.5 inches in an hour. This resulted in another debris flow to make its way towards Flagstaff, which then combined with the ongoing flash flooding within the city. This resulted in cars being carried away by rapidly flowing waters, a dramatic rescue of a father and his two daughters from their car surrounded by rushing flood waters (Fig. 36B-C), and a shelter in place order being issued[16].

The following week, northern CO experienced multiple debris flows and flash flooding. In the evening of July 20th, 30 miles along Poudre Canyon experienced flash flooding and a debris flow associated with the Cameron Peak Fire burn scar[17]. According to the NWS Boulder, about an inch of rain fell in 30 min over the region due to slow moving storms[18]. The Cameron Peak Fire was the second largest fire in CO history. The rain fell over areas with moderate to high soil burn severity, causing debris flows that, due to the terrain, were funneled down into Poudre Canyon and into Poudre River, which Highway 14 runs along. The event closed the highway, washed out bridges, led to mandatory evaluations and unfortunately destroyed 5 structures, one of which had a family in it that sadly perished in the event. Figure 37 shows the location of the rainfall, the flash flood warning that was issued downstream of the heaviest rain, and the aftermath of the event.

---

[15] https://www.weather.gov/fgz/FlagstaffJuly2021

[16] https://abc7.com/arizona-flooding-flash-water-rescue-catalina/10888440/

[17] https://www.cpr.org/2021/07/22/poudre-canyon-flood-search-continues-for-three-missing-people-highway-14-reopens/

[18] https://www.coloradoan.com/story/news/2021/07/23/poudre-river-flooding-explained-colorado-meteorologist-black-hollow-flood/8063039002/

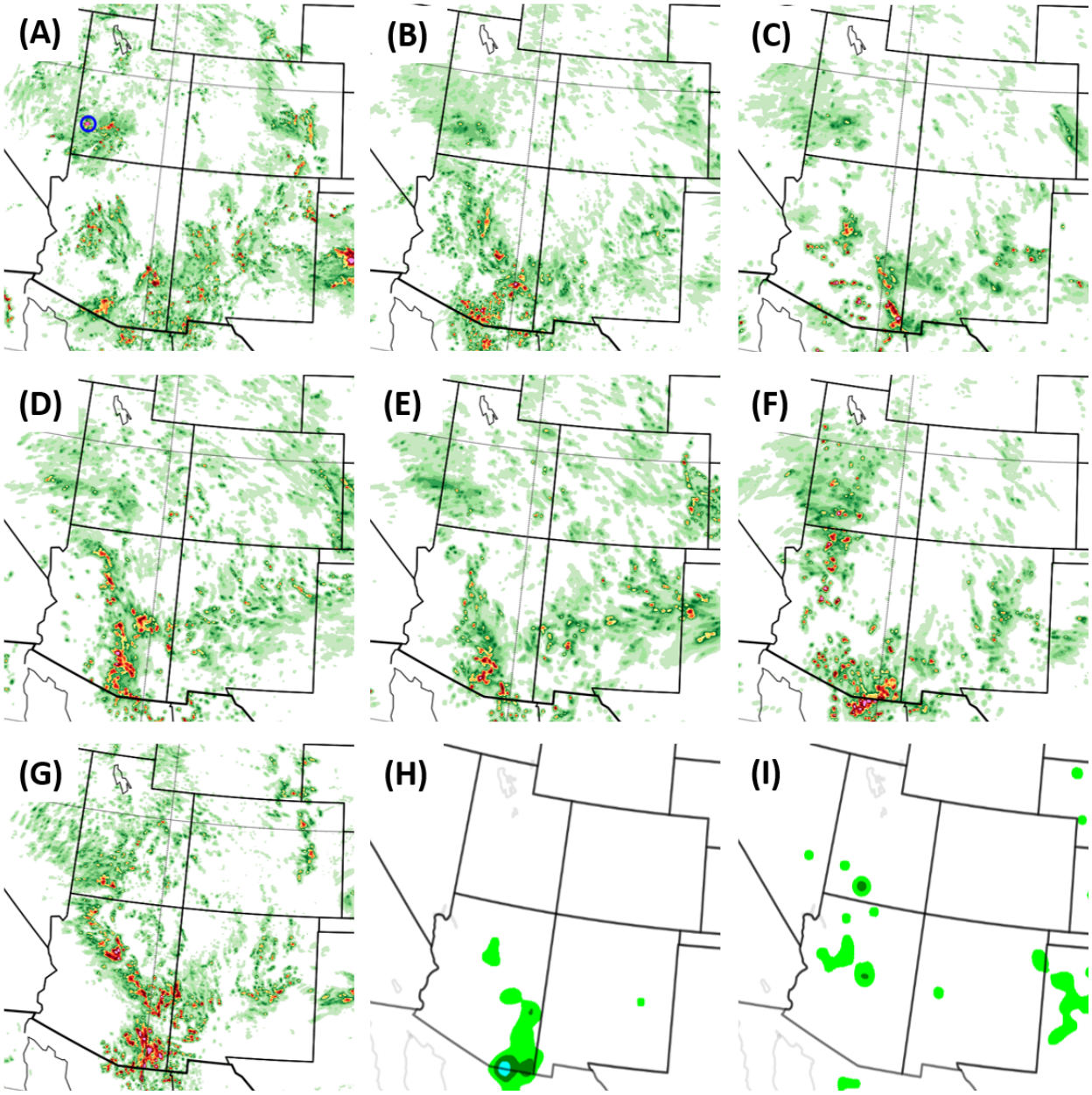


Figure 34: *(A) 6 h MRMS QPE. 6 h QPF from: (B) HRRR, (C) RRFS1, (D) SSEF Cntl, (E) LAMX, (F) LAMDAX, and (G) NAMnest. Probability of exceeding 2 inches in 6 h from (H) HREF and (I) RRFSCE. All valid 21 UTC 15 July to 03 UTC 16 July 2021.*

Figure 35: *(A) The storm total estimated rainfall from around 5:40 AM to 6 PM MST from the KFSX WSR-88D radar. Tweets from Adrian Skabelund: (B) Image of the aftermath from debris flow and (C) still image taken from a video of the debris flow on 13 July 2021. All images courtesy of the Storm Summary written by the Flagstaff WFO.*

Figure 36: *(A) The storm total estimated rainfall from around 8:30 AM to 9 PM MST from the KFSX WSR-88D radar; courtesy of the Storm Summary written by the Flagstaff WFO. (B)Still image taken from a video from ABC7 news in Flagstaff, AZ (ABC7 Staff 2021) of a dad and 2 daughters being rescued from flood waters in Flagstaff. (C) Still image taken from a video from an article written by Bradford C. (2021) of Pruis floating down a street in Flagstaff as flood waters moved through the area.*

Another event during week 4 occurred in the Glenwood Canyon on July 22 due to rain falling on over the Grizzly Creek Fire. The debris flow went across I-70, closing the interstate for multiple days as crews worked to remove debris and check for structural damage on the road. This was not the only time there was a debris flow from this burn scar this summer and closed the interstate. Between June 26-27, multiple rounds of heavy rainfall moved through the area causing numerous mudslides, including one that was 70 feet wide and 5 feet deep in some areas[19]. Then, a week after the July 22 event, on July 29th (after FFaIR ended) over 100 people were stranded on I-70 near Hanging Lake Tunnel overnight as a result of a debris flow[20]; see Fig.

[19] https://www.weather.gov/gjt/GrizzlyCreekDebrisFlows_26-27June2021

[20] https://www.postindependent.com/news/rains-that-triggered-mudslides-in-glenwood-canyon-approached-what-weather-observers-call-a-500-year-event/

38 to see the extent of the damage to the highway. In one location north of Glenwood Canyon 0.52in/15min was recorded and 1.1in/15min was recorded just outside of the burn scar. This was followed two days later by another debris flow in the canyon, with a location south of the canyon receiving 0.7in/15min.

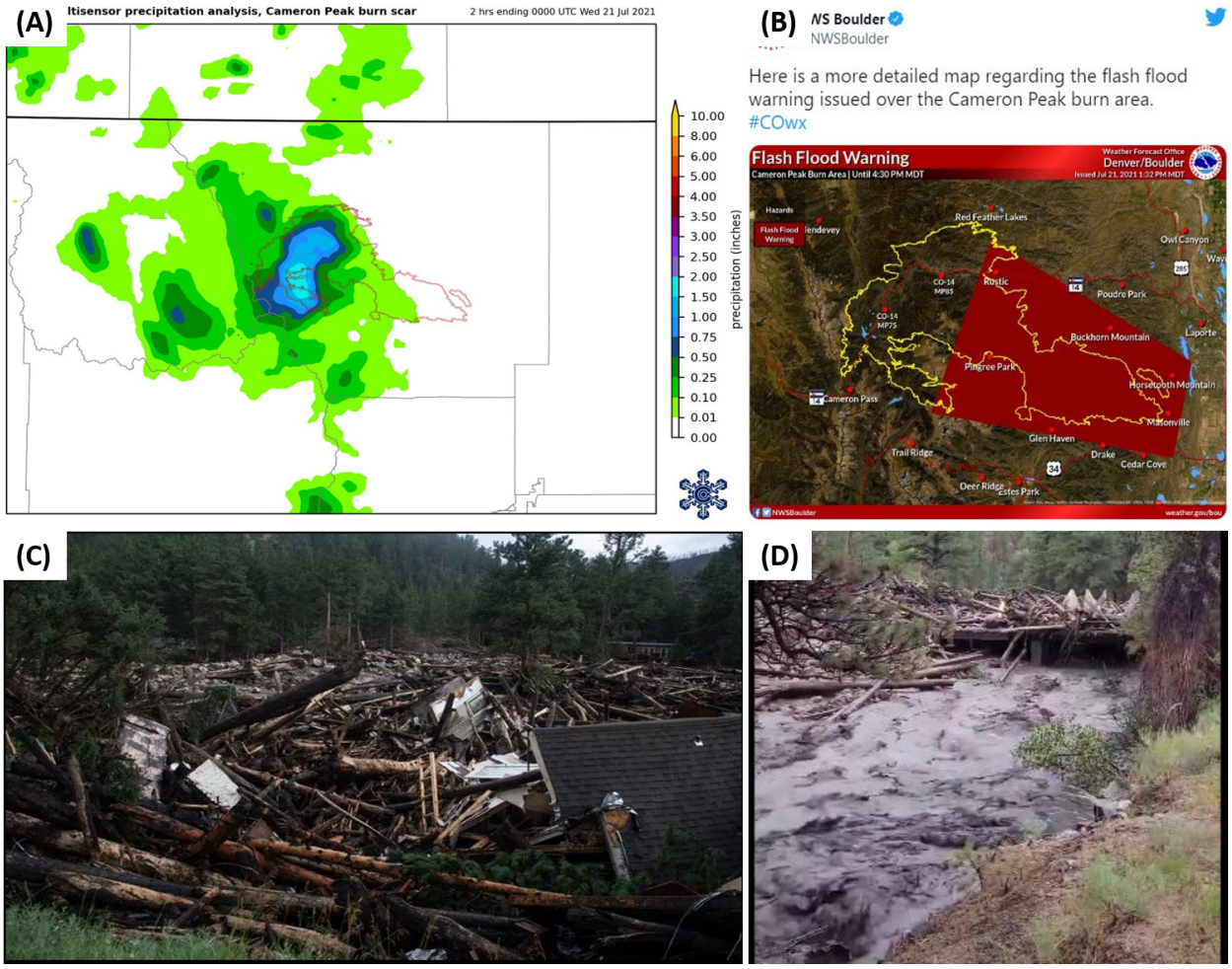


Figure 37: *(A) 2h MRMS QPE valid 00 UTC 21 July 2021. Red contour is the outline of the Cameron Peak burn scar in CO. (B) NWS Boulder's Tweet about the flash flood warning issued for the Cameron Peak burn scar (outlined in yellow). (C) and (D) images of the debris flow that went through Poudre Canyon; courtesy of Blumhardt M. (2021).*

Figure 38: *Images of the damage from debris flow that occurred in Glenwood Canyon on July 22, 2021 due to rain falling on over the Grizzly Creek Fire; courtesy of (left) Meckles and Bianchi (2021) and (right) still taken from drone footage in Stroud, J. (2021).*

The impact of short duration, heavy rainfall on burn scars demonstrates the importance of knowing the possible range of 15 min or shorter rainfall totals so forecasters can proactively communicate the risk of debris flows to the public and our partners. Output from the Warn-on Forecast System (WoFS), which has output every 5 minutes, and the HRRR hourly maximum 15 minute QPF that is produced for FFaIR are some such tools that guide forecasters in this area. When the WoFS was available over the southwest[21], numerous WPC MPDs refenced output from the system to alert partners of the potential of heavy rainfall in short durations over burn scars. For instance, in MPD #538 stated:

> "WoFS ensemble members are increasing with 5 min rain totals of .25 to .35" forecast in numerous members. Ensemble mean rain totals of 1.5" through 05z are expected across N Mohave county, with some 90th percentile members as high as 2.25". This is in general agreement with other Hi-res CAMs including the evolution from the HRRR providing increased confidence toward this evolution, with fairly good placement consistency with each 30 minute run (WoFS)."

Figure 39 shows how important knowing what the maximum rainfall in 15 min can be. Comparing the 6 h QPF from the HRRR to MRMS QPE (Fig. 39A-B), valid 00 UTC to 06 UTC on July 30 , the forecast itself was not very good. It not only had too small of a footprint but also was too light on precipitation totals. This was especially true in the immediate area of the Grizzly Creek burn scar (magenta outline on the images). However by looking at the max rainfall in 15 minutes, a majority of the rainfall forecasted for that time period is accumulated in 15 minutes. For instance, focusing on the pockets of 0.75in/6h forecasted to the northwest of the burn scar, it can be seen that most of that rainfall occurs in a short duration when evaluating the 15 min maximum rainfall at both 02 and 03 UTC. Those types of rates over a burn scar can be enough to create debris flow. Seeing this would suggest to the forecasters that if convection develops

[21] WoFS is still experimental and is not run daily but rather run for specific outlook criteria. WoFS was run a handful of times during FFaIR and when it was available, participants were allowed to look at during their forecast process. Additionally the domain is small, a 900x900 km (https://wof.nssl.noaa.gov/).

slightly to the southwest of where the HRRR is forecasting, rainfall rates would be high enough to cause concern.

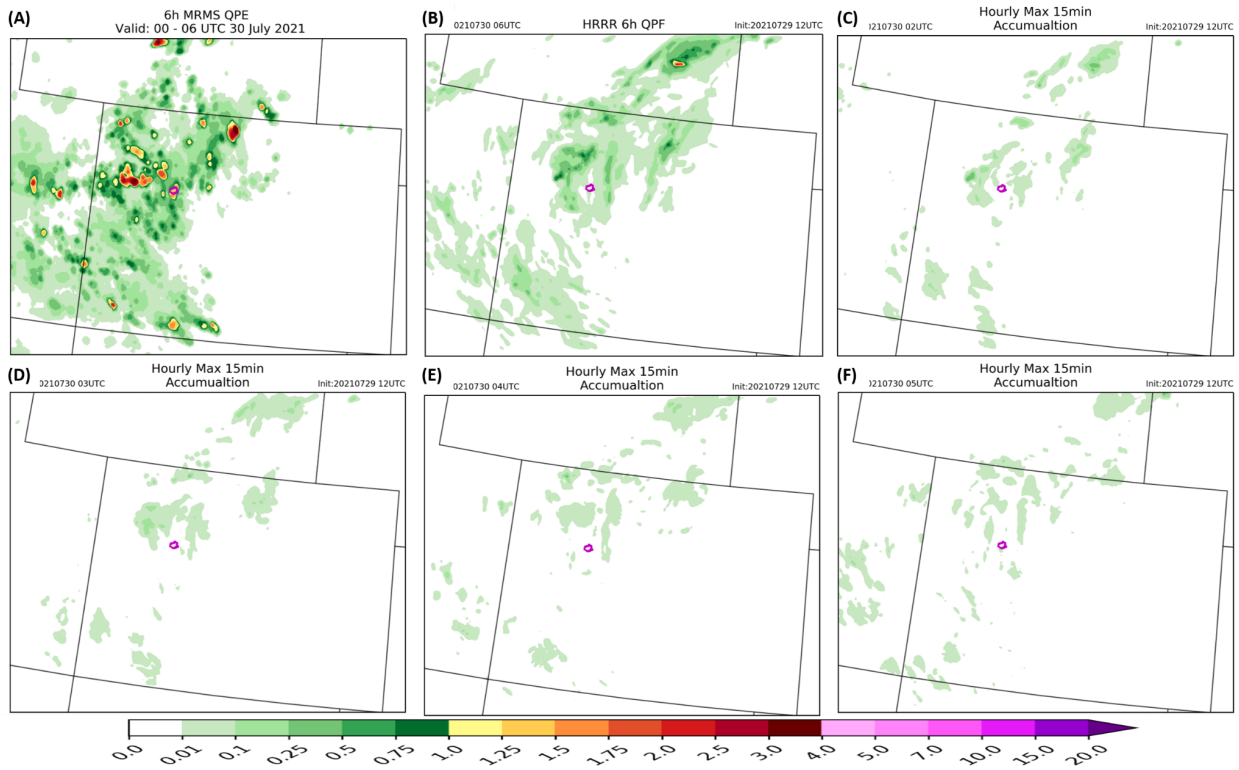


Figure 39: *(A) 6h MRMS QPE and (B) 6h HRRR QPF valid 00 UTC to 06 UTC 30 July 2021. (C)-(F) Hourly maximum 15min QPF from the HRRR valid at (C) 02 UTC, (D) 03 UTC, (E) 04 UTC and (F) 05 UTC 30 July 2021. Location of the Grizzly Creek burn scar outlined in magenta.*

3.2 Flooding Across the World

Although the focus of FFaIR is on the Continental United States, heavy rainfall and flooding events that occur outside of the country are often still of interest to our participants. During FFaIR this year, there were two notable flood events that happened outside the U.S. that were discussed a great deal during the experiment. Since so much time was taken during FFaIR to look at what happened in these events, a brief description is included here.

During week 3, heavy rainfall across eastern Belgium and western Germany resulted in catastrophic flooding and, unfortunately, over 180 deaths across the region[22]. As can be seen in Fig. 40, from July 12th to the 15th parts of western Germany received 100-150 mm (4-6 in) of rain while Belgium saw 100-200 mm (4-7.8 in). Most of this rain occurred on July 14th, with parts of western Germany receiving over 120 mm (4.7 in) and Belgium, Fig. 41, receiving over 85 mm (3.3 in) in 24 hours. According to the Deutscher Wetterdienst (DWD), Germany's meteorological agency, the 24 h totals seen in some areas of Germany were more than twice that

[22] https://www.eumetsat.int/devastating-floods-western-europe

of the monthly average[23]. The enormity of the event was worsened by the fact that the rain fell in the terrain of the eastern mountains of Alps, resulting in floodwaters being funneled down through valleys and into the towns and villages in generally remote areas of Germany. Figure 42 shows some of the destruction from the event.

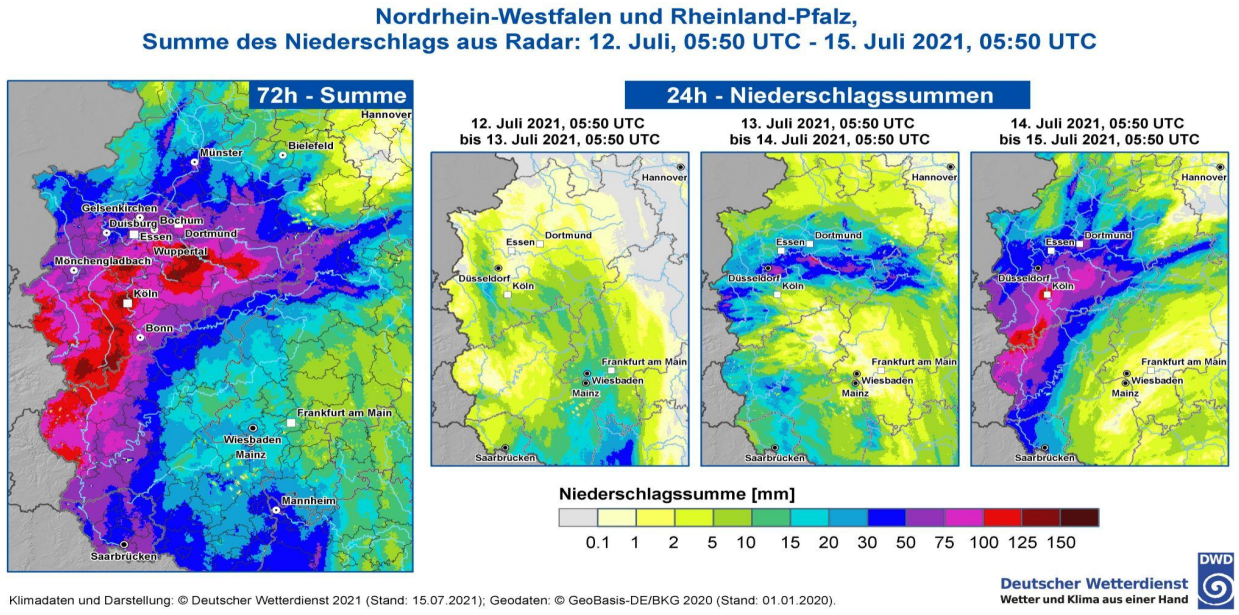


Figure 40: *Radar indicated QPE from the DWD. (Left) 72 h totals valid 0550 UTC 12 July to 0550 UTC 15 July 2021. (Right) 24 h totals from left to right valid: 0550 UTC 12 July to 0550 UTC 13 July 2021, 0550 UTC 13 July to 0550 UTC 14 July 2021, and 0550 UTC 14 July to 0550 UTC 15 July 2021. Courtesy of EUMETSAT article written by Puca et. al (2021).*

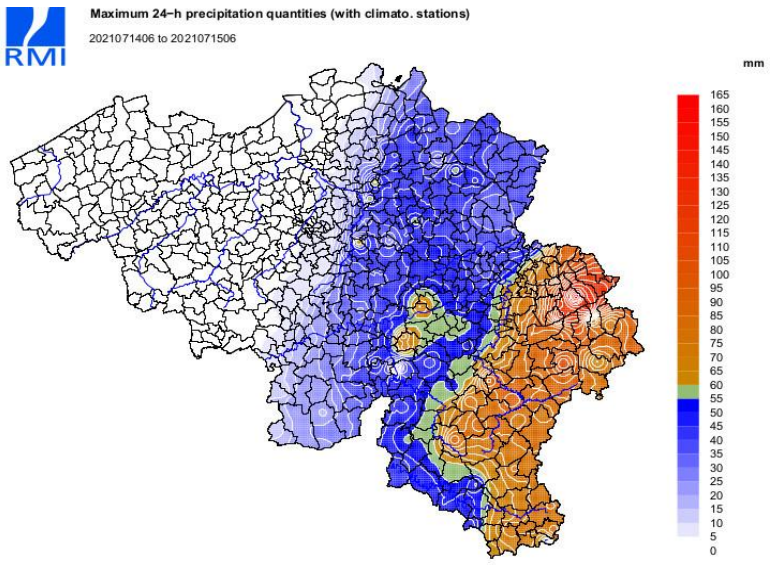


Figure 41: *The maximum 24 h precipitation (mm) in Belgium from their Royal Meteorological Institute (RMI), courtesy of EUMETSAT article written by Puca et. al (2021).*

[23] https://www.washingtonpost.com/world/2021/07/15/germany-flooding-buildings-collapse/

Sadly, although the event was well forecasted by the DWD and an extreme weather alert was issued by the office, as well as by the European Flood Awareness System (EFAS), miscommunication between the local and national levels and insufficient dissemination of the threat led to much of the population being unaware or prepared for the flooding (e.g. Kottasová and Krever 2021[24], Marthiesen, Burchard and Gehrke 2021[25]). Although the failure in communication was devastating, it reiterates the importance of the Decision Support Services (DSS) and outreach with both the public and other government agencies that has been at the forefront of the NWS's mission in recent years. It also shows that having a good forecast is not all that matters, especially if those who need to be aware of the threat do not get the information. As HMT and other testbeds continue to evolve, working with social scientists on how to better communicate forecasts and threats to the public and our partners should be a priority.

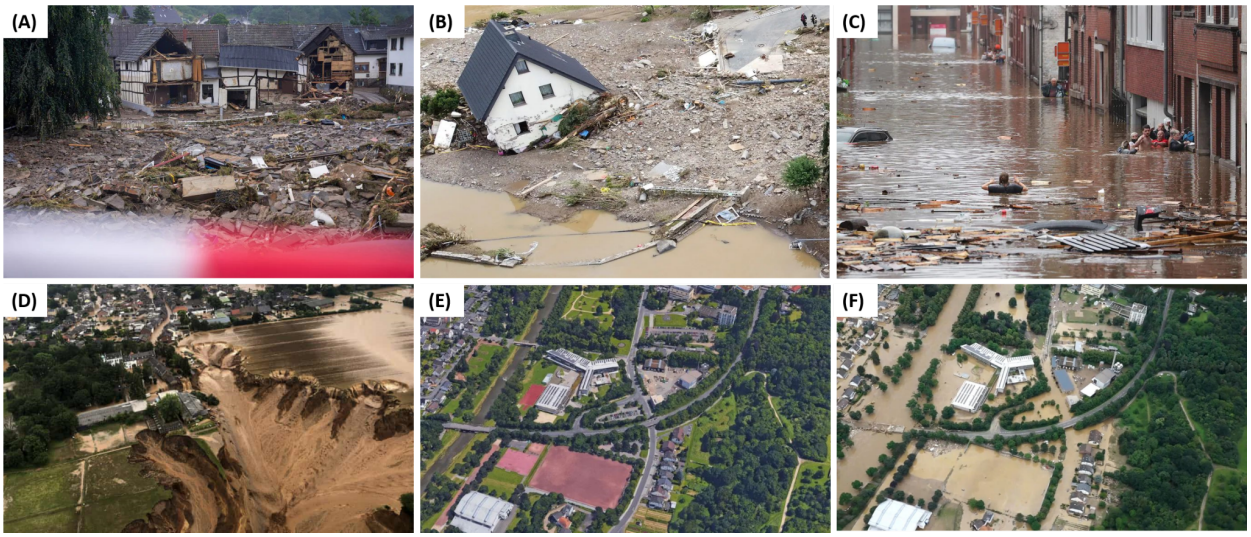


Figure 42: *Images of some of the damage from the flooding in western Germany and Belgium on July 14, 2021. (A) Debris and structural damage and (B) a house pushed off foundation surrounded by debris in Schuld, Germany. (C) People wade in chest deep floodwaters in Liege, Belgium. (D) Landslide from the flooding, North Rhine-Westphalia, Germany. (E) A before and (F) after the flood image of Neuenahr-Ahrweiler, Germany. Images (A)-(C) are courtesy of Washington Post article by Loveday (2021) and (D)-(F) from the Guardian Article (2021)[26].*

The following week, the city of Zhenghou in eastern China experienced an extreme rainfall event that led to devastating and deadly flooding. The 24 h rainfall totals can be seen in Fig. 43. During the 24h period, 195 automated stations statewide recorded rainfall exceeding 250mm and a maximum of 696.9mm (27.44" in 24 hours) was recorded at a dam near Zhengzhou (Zhou, M. MAP email). Furthermore, nearly 8 inches fell in one hour, between 4 and

[24] https://www.cnn.com/2021/07/19/world/netherlands-germany-flood-defense-warning-system-intl-cmd/index.html
[25] https://www.politico.eu/article/germany-floods-dozens-dead-despite-early-warnings/
[26] https://www.theguardian.com/world/2021/jul/16/western-germany-floods-angela-merkel-horror-catastrophe-deaths-missing-search-flooding-belgium

5 pm LST. Flood waters rose quickly across the city, sweeping away cars and bringing the city to a standstill[27]; Fig. 44. This occurred in the midst of rush hour, with thousands of people underground in their subway system as flood waters poured in, trapping passengers in the subway cars. In some instances the water was so high that even standing on the seats the water reached above people's hips[28]; Fig. 44C. In other parts of the city schools had to be evacuated by rescue teams and hospitals lost power leading to patients needing to be transferred to other hospitals[29]. The torrential rainfall and flooding also resulted in a 65 ft breach in a dam located in Louyang City. According to the CNN article written by Nectar Gan and Zixu Wang, at least 33 people lost their lives in the event, with 12 of the deaths occurring in the subway.

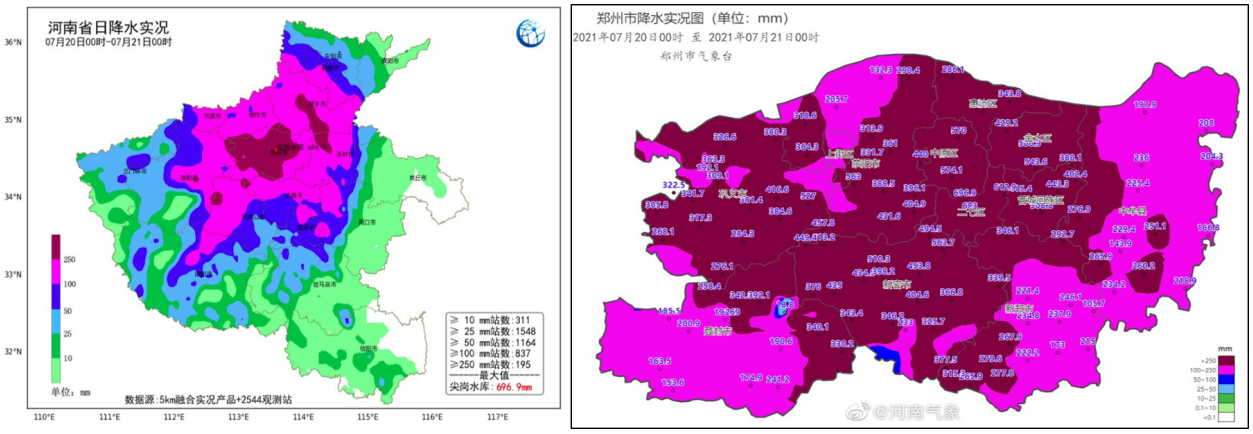


Figure 43: *(Left) the 24-hour total rainfall of Henan Province ending midnight July 21, 2021. The bullseye was centered right over their state capital city, Zhengzhou. (Right) A zoom in over the Zhengzhou metro area. Courtesy of email sent to NOAA MAP from Zhou, M..*

Figure 44: *Images of the aftermath of the extreme rainfall across Zhengzhou, China on July 21, 2021 taken from the article written by Cappucci M. (2021) with (C) being a still image from the video included in the article of a Tweet from Insider Paper.*

[27] https://www.washingtonpost.com/weather/2021/07/21/zhengzhou-china-record-rain-flooding/
[28] https://www.cnn.com/2021/07/22/china/zhengzhou-henan-china-flooding-update-intl-hnk/index.html
[29] https://www.bbc.com/news/world-asia-china-57861067

The region, located along the foothills of the Tai-hang Mountains, is prone to extreme rainfall events since it is affected by the southwest monsoon, easterly trade winds, and occasional troughs and cutoffs from the Northwest during the summer[30]. Figure 45 shows the setup that resulted in the extreme rainfall over Zhengzhou. Two Typhoons over the Pacific, Cempaka and In-fa, led to enhanced low-level moisture and convergence south of the city that resulted in upslope along the mountain range. A coinciding shortwave moving into the region from the west led to upper level divergence over the region. All this occurred amidst the giant monsoon gyre.

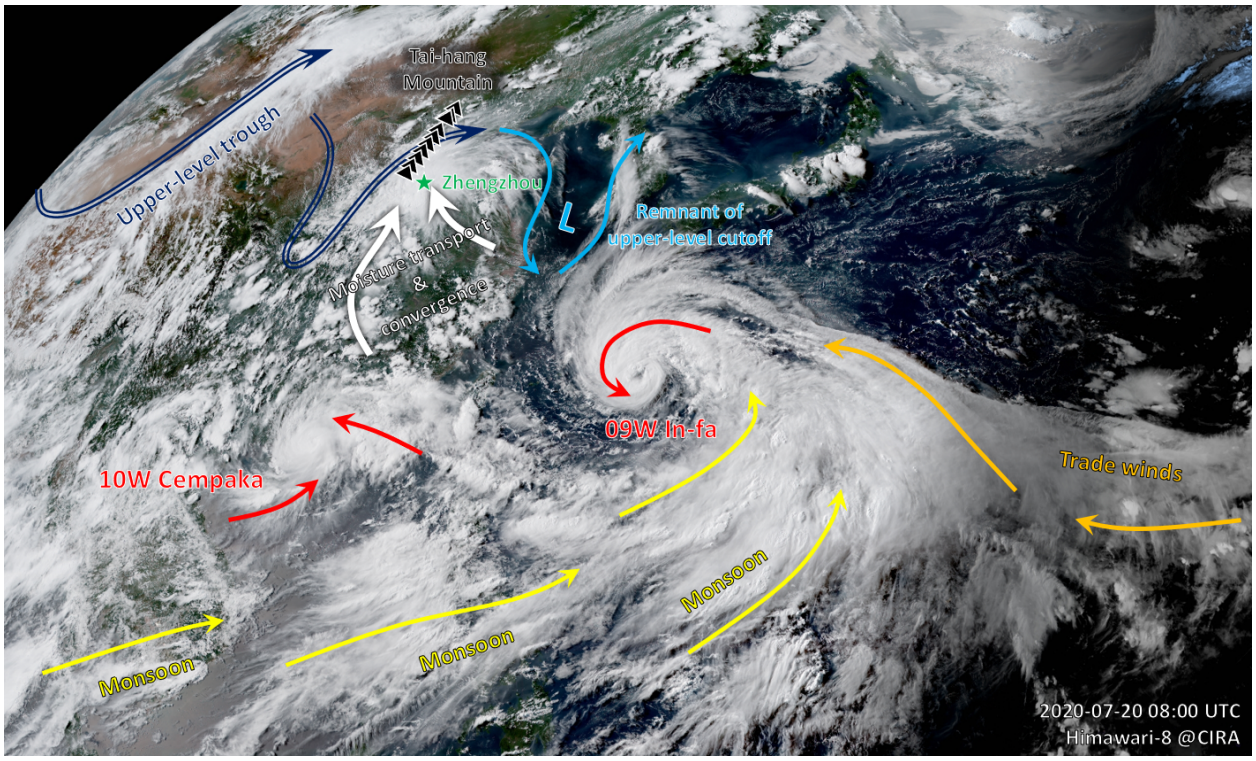


Figure 45: *Image showing the meteorological and topographical setup that led to the flooding across Zhengzhou, China on July 21, 2021. Analysis from the Tweet from Minghao Zhou included in the Washington post article written by Cappucci M. (2021).*

Although it might seem odd to include a section on international flooding events in the FFaIR Report, one of the goals of FFaIR is to provide a space in which participants can discuss all aspects of heavy rainfall and flash flooding. This is not limited to only the United States, since understanding the phenomena that drives these events across the world benefits heavy rainfall and flash flooding forecasts in the United States. Discussion of the difficulties of forecasting these types of events with the international community is important to advancing the science forward. It is also a reason why the HMT should continue to support participants from outside the country. For the past two years we have had a participant from the German Meteorological Service (DWD) and the contributions to the discussions he has provided on events during FFaIR have been thought provoking and insightful. And in the case of the

[30] Information from an email from Minghao Zhou to the NOAA MAP server.

Germany/Belgium flood, he provided additional information on challenges of communicating the threat and information about the climatology of Germany to put the event in perspective. For instance, he explained that they have had similar amounts of rain in eastern Germany with little impact on the region but because the event was in mountainous terrain the impacts were exacerbated. This is similar to what has happened in the Poudre Canyon debris flow and flash flood event, where terrain and location of maximum rainfall worsened the impacts from heavy rainfall, that if had fallen somewhere else, say along the Gulf Coast, would likely not have had such devastating results. Therefore, taking lessons learned from international events can help drive the future of heavy rainfall and flash flooding forecasting and communication of their risk.

4. Results

The qualitative and quantitative (also referred to as subjective and objective) results will be discussed in the following sections. The sections will focus on the results from the deterministic models, the ensembles, the EROs and the MRTPs. Although most of the discussion will be on the analysis of the data and summary of the results, recommendations about the experimental products will also be discussed.

4.1 Deterministic Guidance

Originally eight experimental and two operational models were planned to be evaluated during FFaIR this year. However, there were stability issues with the North American version of the RRFS from GSL (RRFS dev 2) and it was decided to not include it in the experiment. Therefore seven experimental models were evaluated and are listed in Table 3, the two operational models that the experimental models were compared against were the HRRR and NAMnest. Table 3 also provides a basic summary of the model configurations. For a more in-depth summary, please refer to the 2021 FFaIR Operations Plan[31] or Appendix B. Additionally, an error was introduced to the EMC LAMs in the beginning of June that resulted in some of the levels in the lateral boundary conditions (LBC) data being set to the moisture values from the stratosphere. This resulted in a dry bias in the LBCs used for the LAMs. This error was fixed at the end of July and therefore was present during all of FFaIR. However, after looking into the error, EMC found that the impacts of the error were not significant. Additionally, they found that the larger the domain the less of an impact the error had. Therefore, the LAMX, because it is run on the North American domain, was impacted the least by the LBC error. An example of a LAMX run with and without the error can be seen in Fig. 46.

[31] https://docs.google.com/document/d/1aCcQsffKKCcx69YIxAwPThkvUaEvInPIblt9fS88kfM/edit?usp=sharing

Table 3: *Model configuration for the FV3-CAMs provided by EMC, GSL and the four members from the OU-CAPS SSEF that were evaluated deterministically.*

Model	ICs	LBCs	DA	Domain	Micro-physics	LSM
EMC FV3-LAM (LAM)	GFS	GFS	no	CONUS	Thompson	Noah
EMC FV3-LAMX (LAMX)	GFS	GFS	no	North American	Thompson	Noah
EMC FV3-LAMDAX (LAMDAX)	Own	GFS	yes	CONUS	Thompson	Noah
GSL RRFS-dev 1 (RRFS1)	Cycled	13km FV3LAM	yes	CONUS	Thompson	RUC
OU-CAPS M0B0L0 (SSEF Cntl member)	GFS	GFS	no	CONUS	Thompson	Noah
OU-CAPS M0B2L1 (SSEF RRFS-like member)	GFS	GFS	no	CONUS	Thompson	Noah-MP
OU-CAPS M0B0L2 (SSEF HRRR-like member)	GFS	GFS	No	CONUS	Thompson	RUC
OU-CAPS M1B0L0 (SSEF WoFS-like member)	GEFS	GEFS	No	CONUS	NSSL	Noah

4.1.1 24 h QPF

As discussed in Section 2.3.1, participants were asked to evaluate model 24 h QPF compared to the observed rainfall and score it from 1 (poor) to 10 (great). Since most of the models were experimental, the number of days in which their forecasts were scored varied. The top image in both Figs. 47 and 48 show the number of days each model was available to be scored by the participants. For both the 00z and 12z cycles the participants felt HRRR performed the best, with an average score of 6.78 and 6.45 respectively. It also was the most likely model to receive a score of 7 or better for each model cycle, with 62% and 52% of the scores being 7 or higher (bottom image in Figs. 47 and 48). This follows the results from last year, where what was then referred to as the HRRRv4 (the now operational version of the HRRR) had the highest average score subjectively. The NAMnest, LAM, and LAMX all performed comparable to one another. For the 00z (12z) cycle the average scores were 6.13 (6.08), 6.06, (6.15), and 6.06 (6.15) respectively. Meanwhile, the 00z LAMDAX and all the SSEF members[32] performed similar to one another. The HRRR-like member had the greatest average of this group with 5.62, followed by the LAMDAX (5.59), the RRFS-like member (5.44), the WoFS-like member (5.4) and then the CNTL member (5.39). The 00z RRFS1 was the worst performer (subjectively) with an average of 5.02. At 12z both the LAMDAX and RRFS1 average scores increased to 5.95 and 5.63 respectively.

[32] The SSEF members were only available at 00z.

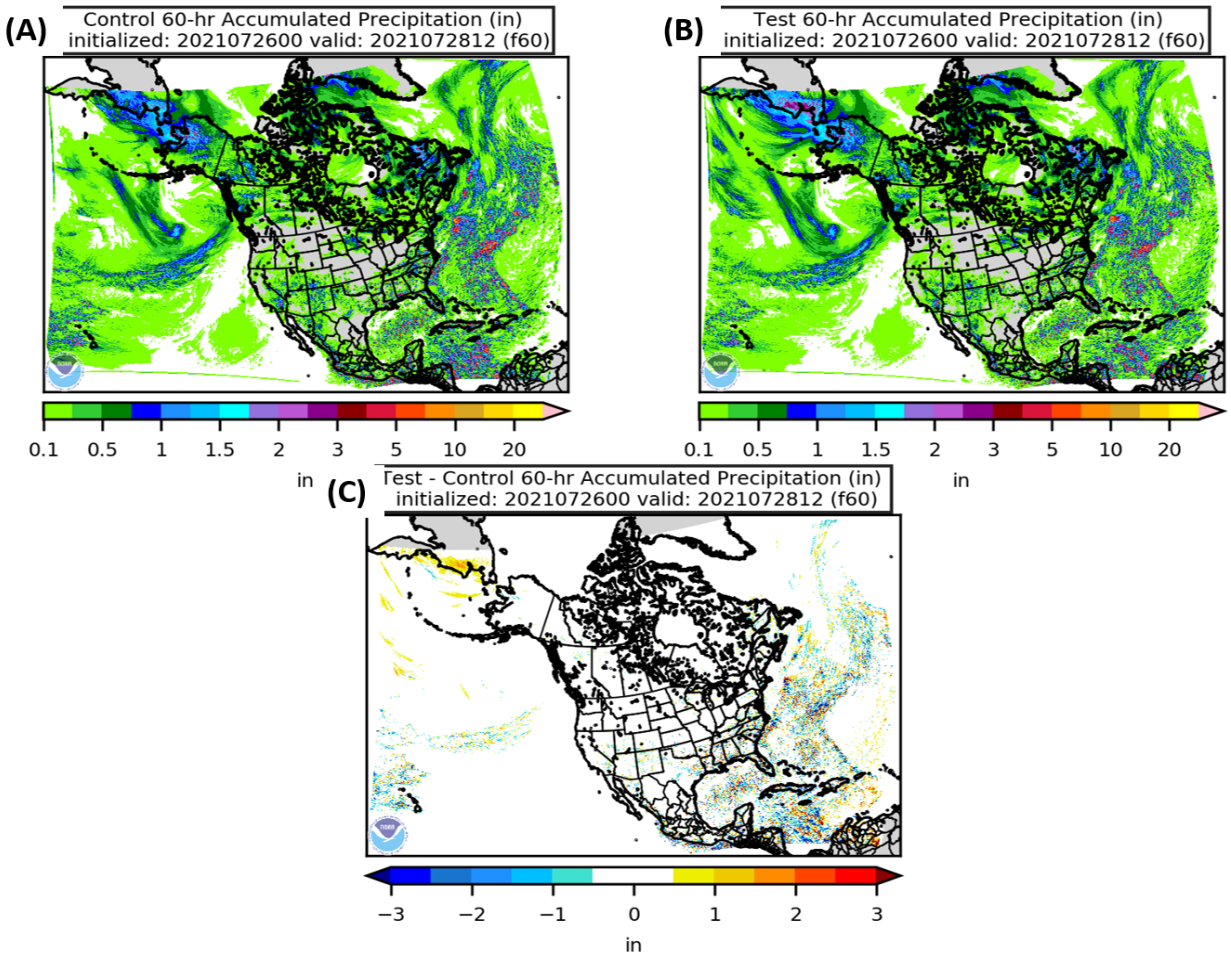


Figure 46: *Plot of 60 h LAMX accumulated QPF (A) with the LBC issue and (B) without the LBC issue. (C) The difference between (B) and (A). All valid at 12 UTC 28 July 2021. Image provided by Jacob Carley from EMC.*

The similarity of the LAM and LAMX to one another and the SSEF members to one another was often noted in the comments from the participants. For instance, for verification on July 1st, one participant noted "LAM did pretty well with location of higher amounts. .... LAMX nearly similar" while the following day another noted "LAMDAX poorest of three LAMs. I think the LAM and LAMX were nearly identical." Comments like these about the LAM and LAMX, as well as having nearly the same distribution of scores (see Fig. 47 and Fig. 48), is encouraging, as the only difference between the two models was the domain size. Since the goal of the NWS is to run the new RRFS system[33] on the larger North American domain, it is essential that there are no major differences between these two models; the objective results also showed little difference between the two and will be discussed later. As for the SSEF, they were less often mentioned in the comments written by the participants, though when they were mentioned it was often discussing the lack of differences among them. However one common theme for

[33] The term FV3 LAM will eventually go away and the FV3 CAM model and ensemble system will be referred to as the RRFS.

the SSEF members was displacement issues, with participants often noting the footprint was ok but it was shifted north/south/east/west. This was likely the result of the SSEF being highly dependent on the GFS forecast; refer to Figs . 17 and the discussion surrounding the images for an example of how impactful the initial conditions from the GFS were in the SSEF at longer lead times.

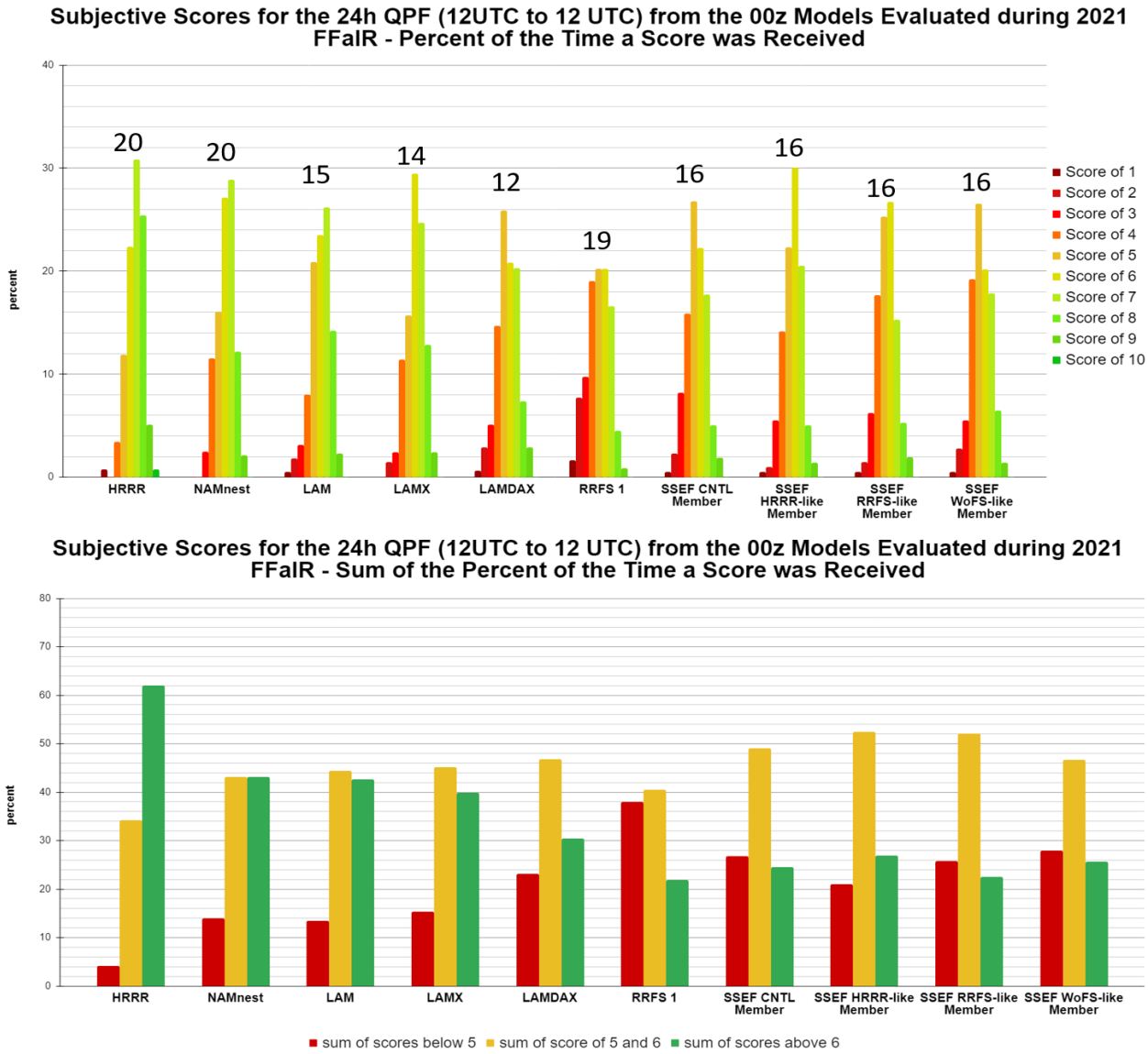


Figure 47: *Results from the subjective verification for 24 h QPF for 00z model initialization showing (Top) the percentage of the time each deterministic model received a score of 1 through 10 during the course of the experiment along with the experimental average plotted above the percentage analysis and (Bottom) the summed percent of the time that each model received a score from 1 to 4 (red), 5 and 6 (yellow) and from 7 to 10 (green).*

Figure 48: *Same as Fig. 47 but for the subjective analysis 24h QPF from the 12z cycles.*

Comparing the 00z runs to the 12z runs, participants overall felt that the experimental models' performance were better in the later run whereas the operational runs were thought to be slightly worse at the later cycle. The RRFS1 saw the greatest increase in average score, from 5.02 to 5.63. This increase in "goodness" between 00z and 12z can be clearly seen in the percent of scores below 5 (bottom images in Fig. 47 and 48), with 38% of the scores being below 5 at 00z compared to 20% for the 12z cycle. Additionally, comments from the participants were full of statements about the "jump in the goodness" of the 12z RRFS1 compared to the 00z RRFS1. For instance, for the forecast valid at 12 UTC July 2 comments like "rrfs1 more popcorn, but less than 00z" and "RRFS1 finally resembles the event that occurred. Location and amounts are still off, but much improved over the 00z run" were made. The difference between the 00z and 12z forecasts from the LAMX are shown in Fig. 49D-E, note that although there are differences

between the two runs, the general pattern and location of precipitation is similar. The RRFS1 on the other hand (Fig. 49B-C), has drastically different solutions between the two runs, with the 00z run incorrectly forecasting the dominant storm mode (showing widespread convection rather than frontal rainfall). This is especially troublesome since the rainfall this day was driven by a well defined, slow moving front (see Fig. 15). The FFaIR team advises GSL to look further into the cause of the sometimes extreme differences between the 00z and 12z runs of the RRFS1. It is possible that the method used in cycling for the ICs or the use of the 13km LAM could be the culprit.

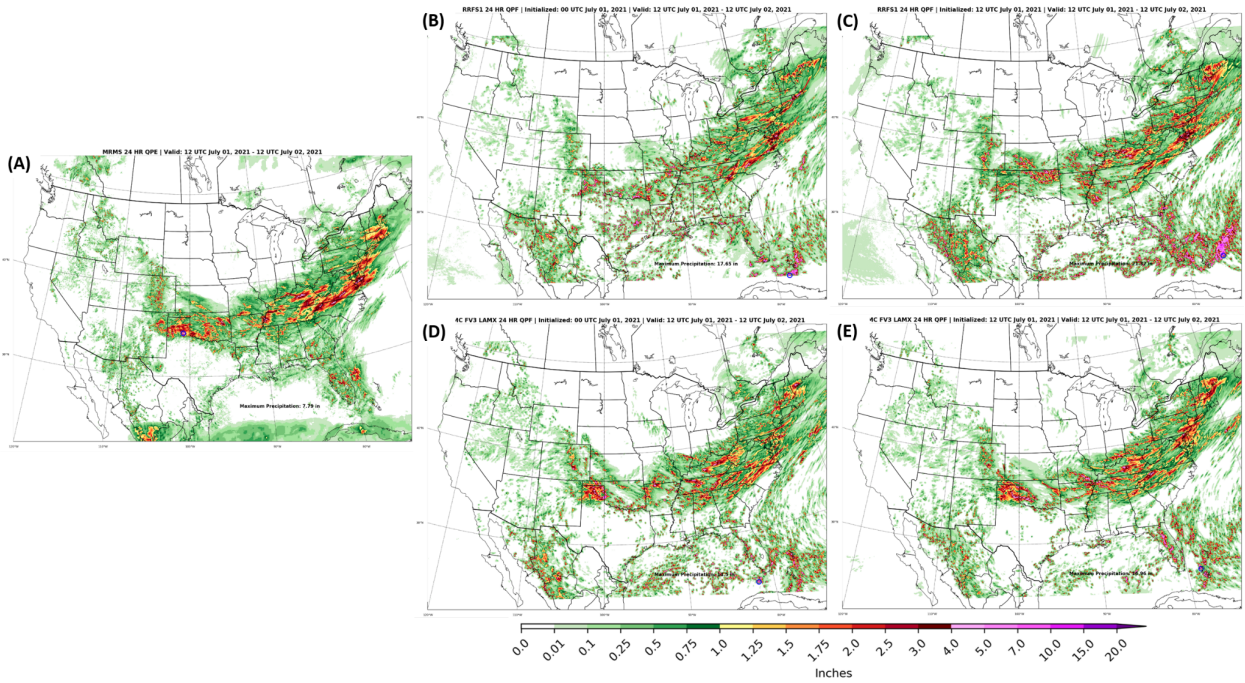


Figure 49: *(A) 24h MRMS QPE. 24h QPF from the (B) 00z run and (C) 12z run of the RRFS1 and (D) 00z run and (E) 12z run of the LAMX. All valid 12 UTC 01 July to 12 UTC 02 July 2020.*

One of the findings in the 2020 FFaIR Experiment was the over-development single cell convection, referred to as popcorn convection by the FV3-CAMs, as well as an overall wet bias. This over-development was not just in the number of cells simulated but also in the QPF totals, often on an hourly timescale. In other words, nearly every cell had precipitation totals noticeably greater than what would be expected out of normal scattered convection, i.e. 2+ inches forecasted in nearly every cell developed. Additionally, the cells looked grid-like in nature; see Figs. 28 and 29 in the 2020 FFaIR Final Report[34] for reference. This year in FFaIR, although there was no formal question asked during evaluation focusing on the FV3-CAMs simulation of popcorn convection or any possible wet bias, comments about the issues were made in both the written comments and in open discussions about model performance. Overall, participants who looked at the FV3-CAMs last year (then referred to as SARs rather than LAMs or RRSF) stated

[34] Final Report can be found here:
https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf

that when it came to the EMC LAMs the over-development of popcorn convection was less pronounced, though still noticeable, than last year. They also noted a wet bias in the LAMs, though again, this appeared to be subdued compared to last year. As for the RRFS1, almost daily the participants commented on the excessive QPF values that were being forecasted by the model as well as the widespread, high precipitation popcorn storms. This problem was exacerbated by the fact that often, especially when looking at the 00z cycle, the forecasted pattern did not appear correct. The SSEF members, though not commented on as much by the participants, also had the tendency to over produce popcorn convection.

An example of the over forecasting of popcorn convection in the FV3-CAMs can be seen in Figs. 50 and 51 for July 15, 2021. This day was particularly active, with three distinct areas of concern. The first was the MCS that moved across KS/MO/IL, the second was the monsoonal convection across the southwest, and the third, which will be the focus of discussion, was the convection over the Gulf states. Comparing the 00z forecasts from all the models, it can be seen that none of them truly captured the evolution of the Gulf convection, which had widespread pockets of 24 h rainfall totals between 2 and 3 inches. Instead, all models forecasted the diurnally driven (popcorn) convection across the Gulf states, with scattered, single cell convection rainfall. However, there is a clear difference between the look of the single cell convective storms in the operational models (Figs. 50B-C) compared to the experimental models. The cells developed in the HRRR and NAMnest appear smaller and generally forecasted rainfall totals from the storms around an 1 inch. Opposing this, the experimental models' cells appear larger and more often than not forecast rainfall totals exceeding 1.75 inches. Even when looking at the forecasted and observed hourly totals across this region, as can be seen in Fig. 52, the larger size of the cells and the higher totals seen in the FV3-CAMs are present. A majority of the cells from the LAM have hourly rates exceeding 2 inches and from the RRFS1 exceeding 3 inches. Lastly, both the LAM and RRFS1 hourly QPF max was within one of these cells (6.01" and 9.14" respectively) while the hourly max for the MRMS and HRRR were not located across the southeast where the popcorn convection was occuring (4.2" in AZ and 3.52" in KS/MO respectively).

As stated, the apparent size of the convection (and likely the size of the updraft) is larger in the experimental models. The overwhelming majority of the cells from the FV3-CAMs forecast over 1.75 inches of precipitation compared to the operational models. The exceptions to this were the SSEF CNTL and RRFS-like members (Figs. 51C and E), which had a QPF pattern more similar to the HRRR and NAMnest. Additionally, the convective cells appear to be smaller in these two FV3-CAMs than in the other FV3-CAMs. The smaller cells and lower QPF from the RRFS-like and CNTL members seems to be a consistent trend throughout the experiment based on a quick, subjective analysis of all days with popcorn convection. However, unlike the other FV3-CAMs, the RRFS-like member consistently produced widespread light rain (<0.1 inch) over

the oceans, similar to what is done by the NAMnest[35]. An additional example of the broad brush of QPF < 0.1 inch produced by the NAMnest and the RRFS-like member compared to the other SSEF members can be seen in Fig. 53.

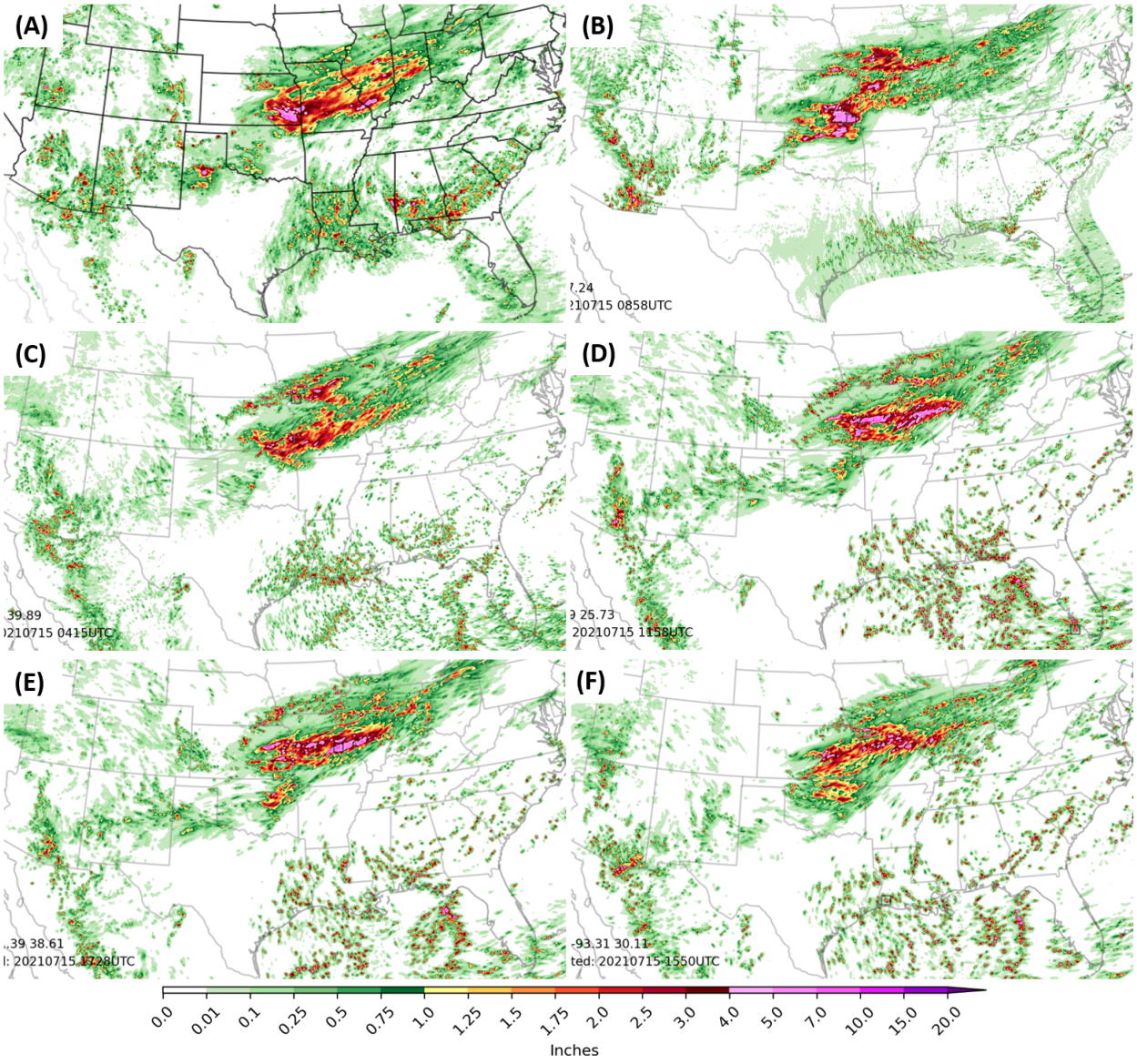


Figure 50: *24 h (A) MRMS QPE and (B) NAMnest, (C) HRRR, (D) LAM, (E) LAMX, and (F) LAMDAX QPF valid 12 UTC 15 July to 12 UTC 16 July 2021.*

[35] The excessive over forecasting of QPF > 0.10 in by the NAMnest over oceans resulted in long computing times to plot the NAMnest QPF images, therefore a mask was nearly always applied to the plots so that only QPF over/near the CONUS would be plotted. Therefore this known issue in the NAMnest will not be seen in most FFaIR plots.

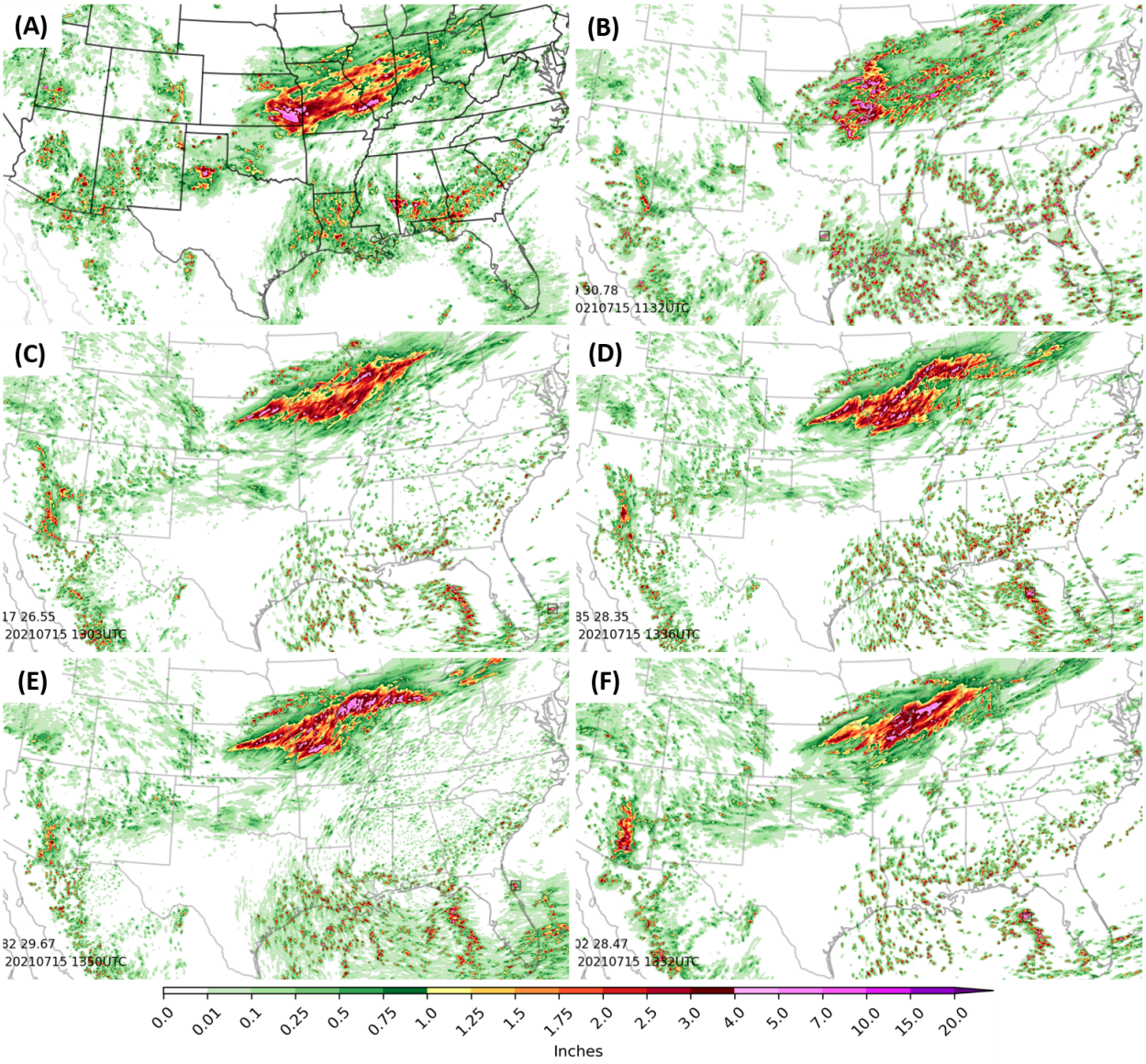


Figure 51: *Same as Fig. 50 but 24 h QPF from (B) RRFS1, (C) SSEF CNTL member, (D) HRRR-like member, (E) RRFS-like member and (F) WoFS-like member.*

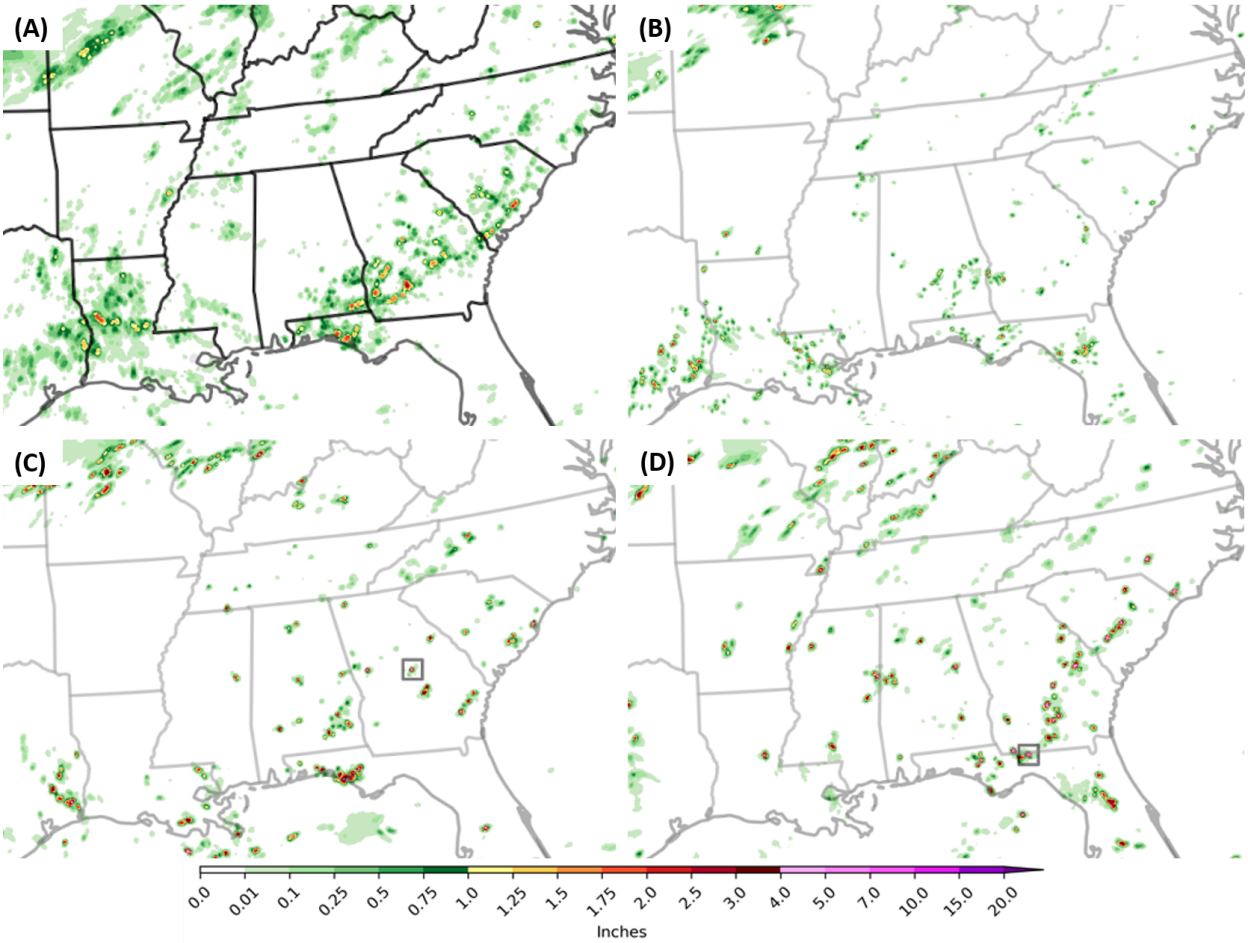


Figure 52: *1h (A) MRMS QPE and (B) HRRR, (C) LAM, and (D) RRFS1 QPF valid 21 UTC 15 July 2021. The grey box indicates the location of the model maximum across the CONUS; neither the MRMS or HRRR CONUS maximum was located in the southeast but both LAM (6.01") and RRFS1 (9.14") were.*

It can be seen in the performance diagrams (hereafter PD) for both the 00z (Fig. 54) and 12z (Figs. 55-56) cycles, that the wet bias noted by the participants was not seen in the low-end QPF amounts, specifically at the half inch threshold. In fact, for the 00z runs, all models had a dry bias at a half inch. Even for the 1 inch threshold, the wet bias from the FV3-CAMs is only slight and is comparable to the NAMnest. However, the PDs at QPF thresholds of 2 inches or greater show a clear wet bias from the FV3-CAMs, that increases with each threshold at a great rate than the wet bias seen in the NAMnest. For instance, for the 12z cycles, the RRFS1 has a bias of roughly 1.25, 2.5, and 5 for the 1, 2, and 3 inch thresholds, while the NAMnest bias for those same threshold stayed between 1.25 and 2.5 as the threshold increased.

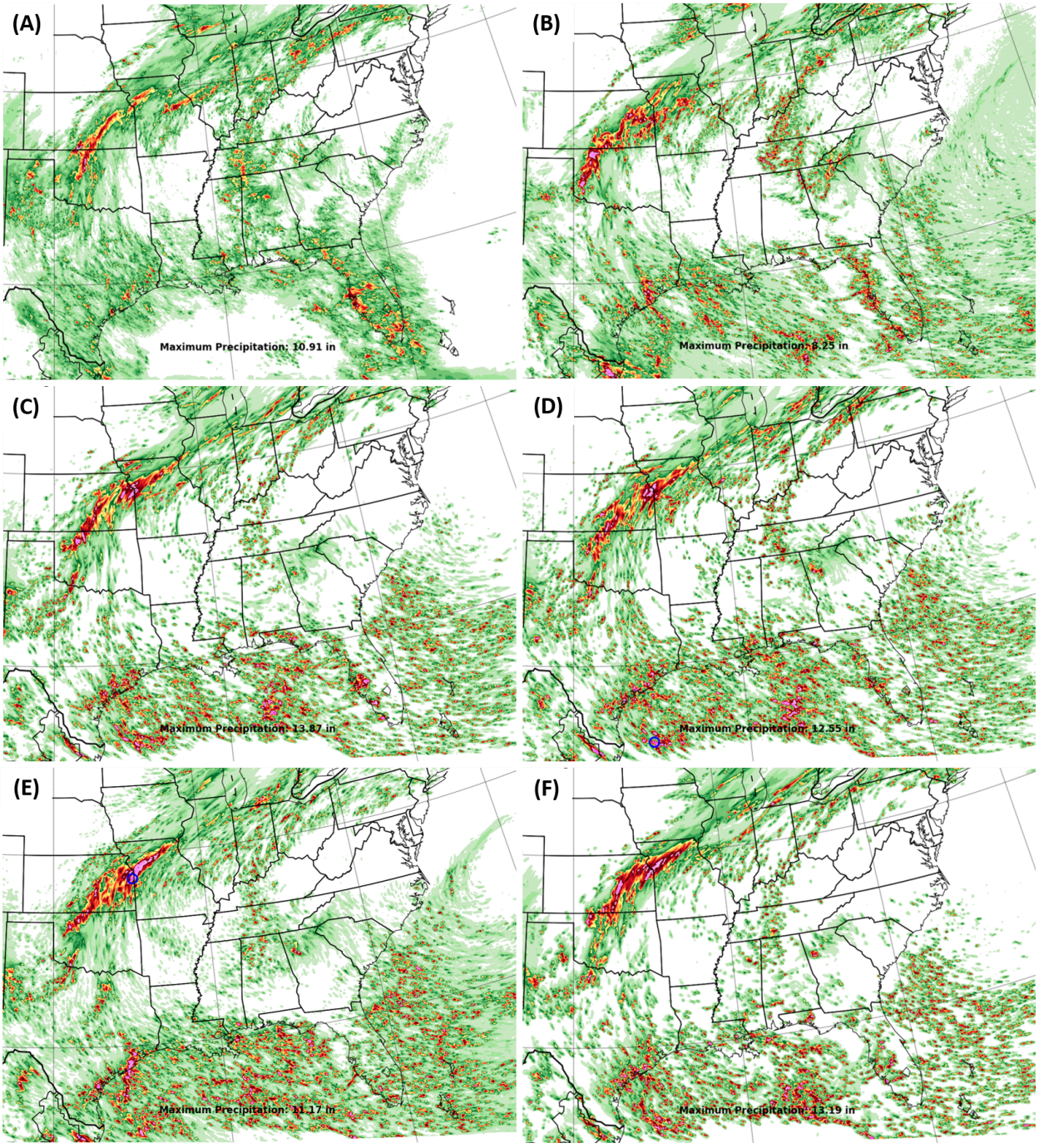


Figure 53: *24 h (A) MRMS QPE and (B) NAMnest, (C) SSEF CNTL member, (D) HRRR-like member, (E) RRFS-like member and (F) WoFS-like member valid 12 UTC 29 June to 12 UTC 30 June 2021.*

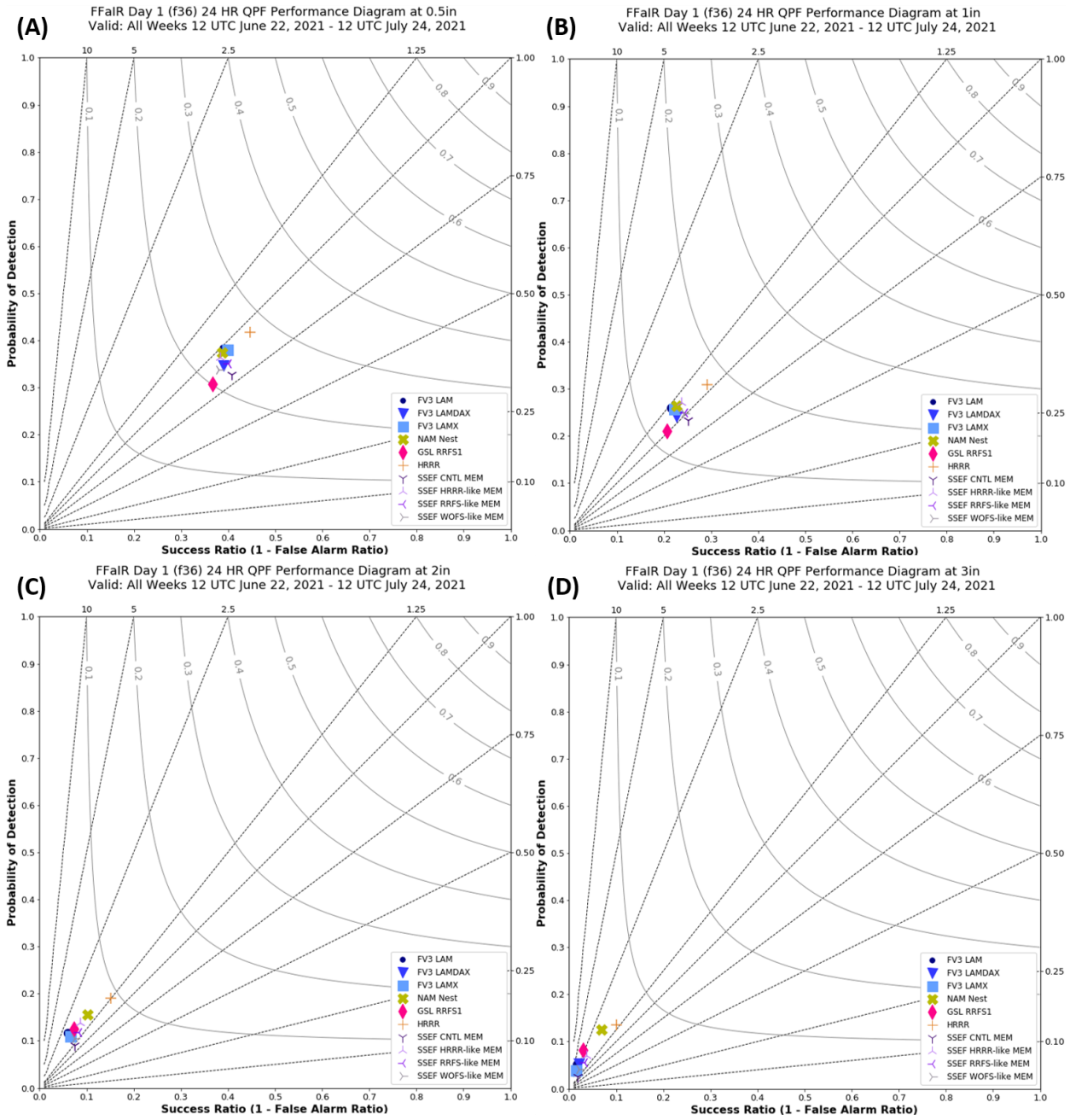


Figure 54: *Performance diagrams for the 00z QPF forecasts valid for Day 1 for only the days in which FFaIR was in session, from June 22 to July 23, 2021 for the deterministic models evaluated during FFaIR. Precipitation thresholds are for: (A) 0.5 inches, (B) 1 inch, (C) 2 inches and (D) 3 inches. Models provided by the same collaborators are in shades of the same color, i.e. EMC models are in blue and SSEF members are in purple.*

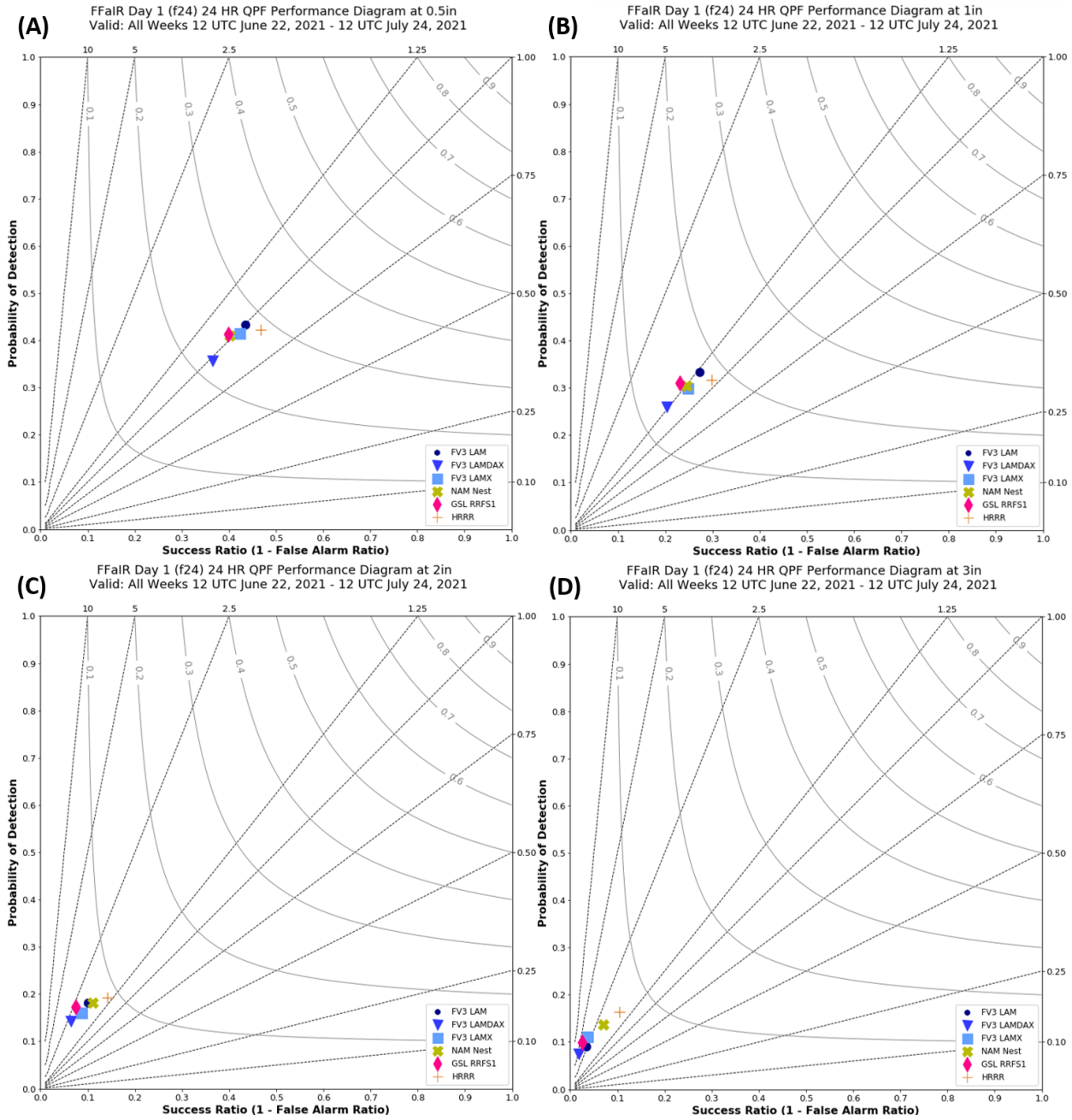


Figure 55: *Same as Fig. 54 but for the 12z cycles.*

The transition from a dry (near-zero) bias at low thresholds to a high wet bias in the FV3-CAMs suggests that the models might be under forecasting precipitation area but over forecasting magnitude. Comparison of weekly and FFaIR long QPF/QPE totals allude to this likelihood. For instance, Fig. 57 shows the weekly QPF/QPE totals during Week 2 of FFaIR. As discussed in Section 3, this week's rainfall was driven by a slow moving front across the Eastern/Central portions of the CONUS. This resulted in an observed, wide, relatively coherent swath of 1+ inches of rainfall from NM, through MO into the Ohio Valley and across New England (Fig. 57A). However, the QPF totals for that week from the LAMX and RRFS1 (Figs.

57C-D) show an inconsistent swath of precipitation over the same area. But where there is rainfall forecasted, the overall QPF totals are greater than or equal to 4 inches when only a few regions of 4+ inches were observed (e.g. OK into the TX Panhandle). The HRRR (Fig. 57B) also tended to over forecast totals across that region. However when focusing from TX to the Carolinas, aside from along the TX/LA coastline, there was a general under forecasting of rainfall totals. Opposing this, a multitude of 4+ inch totals are speckled across the region in the LAMX and to a more notable extent the RRFS1 over the Gulf states. Furthermore, like with the previous area of concern, across the south no coherent QPF footprint like the ones observed were seen. The look of the QPF footprint and the amounts also show the impact the forecasting of popcorn convection had across the south in the FV3-CAMs.

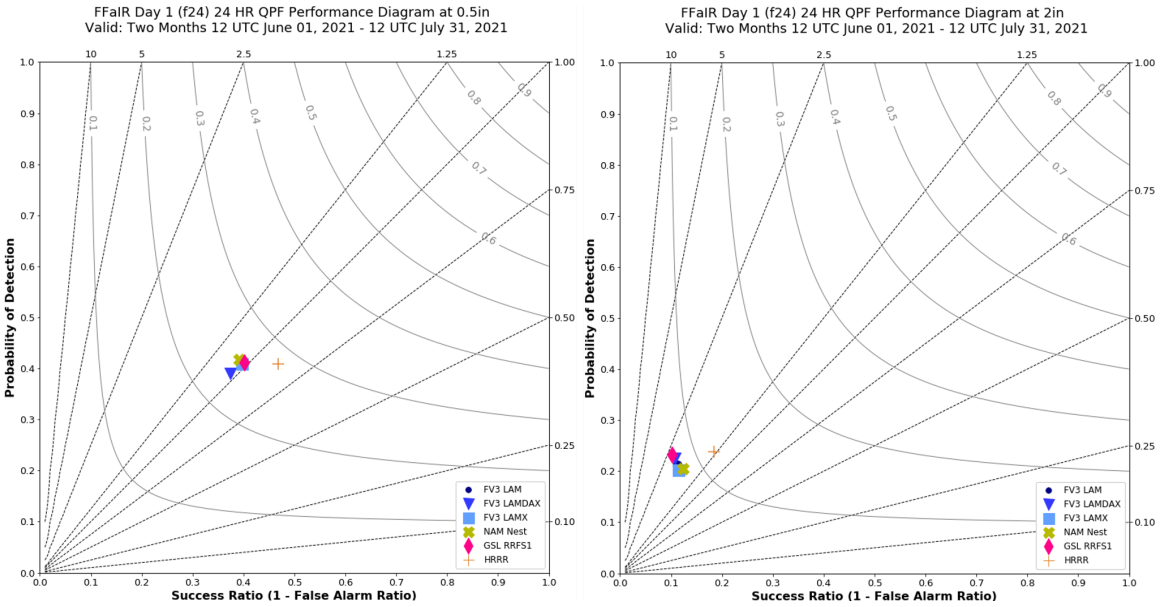


Figure 56: *Performance diagrams for the 12z QPF forecasts valid for Day 1 from June 01 to July 31, 2021 for the deterministic models evaluated during FFaIR. Precipitation thresholds are for: (A) 0.5 inches and (B) 2 inches.*

Further support of the hypothesis that the FV3-CAMs might be underpredicting the rainfall footprint but over predicting totals where it does rain can be seen in Figs. 58-60 ; these show the 00z cycles of LAMX, RRFS1, and HRRR 6h QPF grid point sums compared to MRMS[36]. Looking at the LAMX (Fig. 59) and RRFS1 (Fig. 60) it can be seen that around the 2 inch threshold the number of model grid points of 2+ inches begins to overtake the number of MRMS grid points for the threshold[37]. However, the HRRR (Fig. 58) does not have the same jump. In fact, aside from the apparent timing issue (seen by the shift in the drop off in high end values compared to MRMS between 18 and 00 UTC), the HRRR slightly under forecasts high end QPF totals. This over forecasting of high end amounts is particularly true for the RRFS1,

[36] MRMS is on a 1km grid. This was upscaled to the 3km grid used by the models. The upscale method used preserves the max values.
[37] Note that starting at 2+ inches the grid count scale begins to differ among the models

especially in the first 12 hours of the forecast, where for the 3 inch threshold, the RRFS1 has over 2,500 more grid points than MRMS six hours into the forecast. The same over forecast is not seen in the HRRR or the LAMX. Interestingly for the lower end amounts (<1 inch) the RRFS1 is closer to MRMS in the first 12 hours of the forecast compared to the HRRR and LAMX.

Focusing on the diurnal cycle in the 6h QPF occurrence plots in Figs. 58-60 at thresholds of 3+ inches, MRMS data at these thresholds show a relative minimum in spatial extent around 00 UTC. However both the RRFS1 and LAMX peak around 00 UTC with coverage of higher end precipitation totals; the HRRR does not show the same trend. This peak corresponds with daytime heating, which often initiates the development of popcorn convection in the warm season. Therefore the over forecasting of rainfall totals associated with this type of convection is likely driving this spike around 00 UTC in the LAMX and RRFS1.

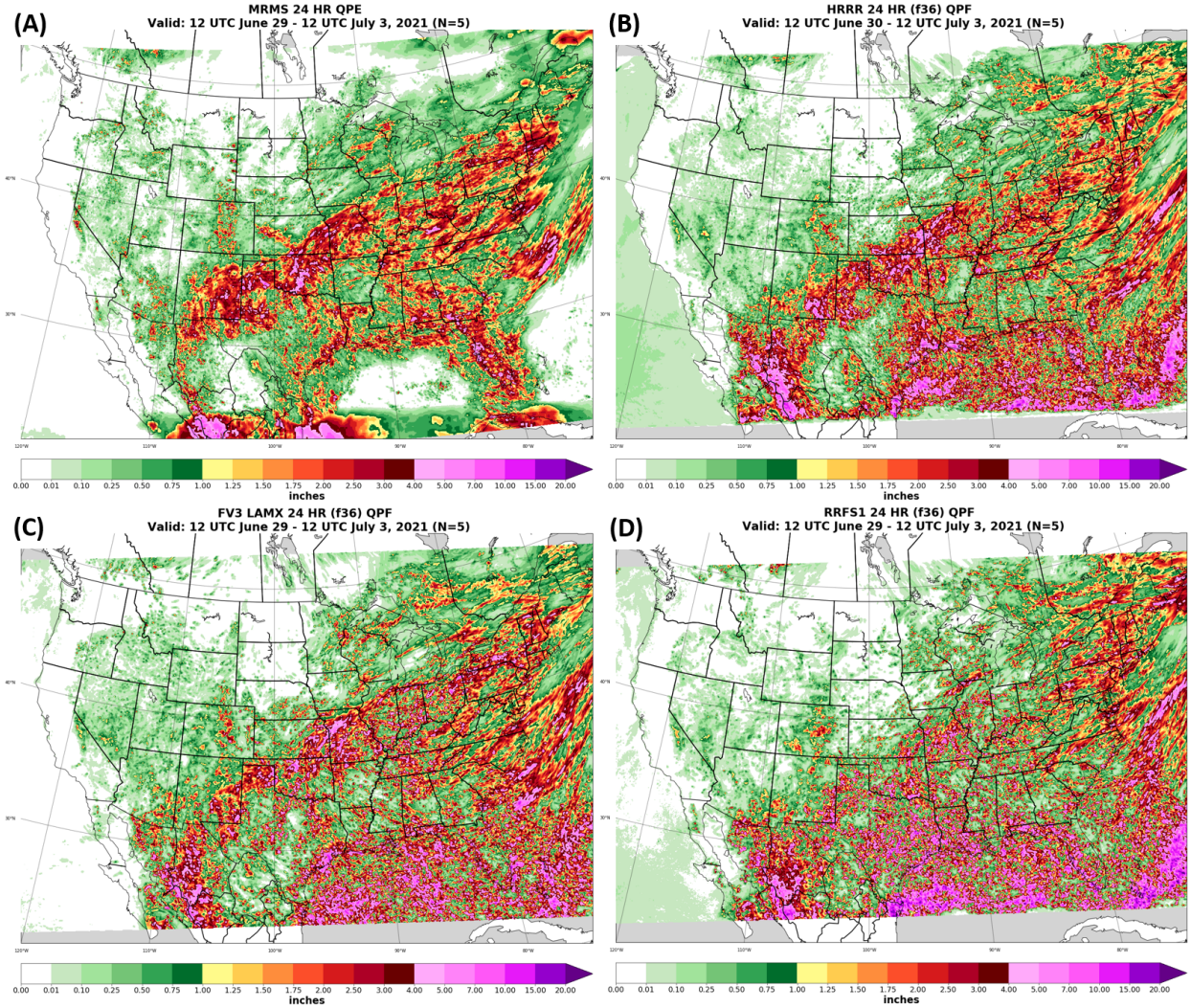


Figure 57: *Accumulated 24h (A) MRMS QPE and (B) HRRR, (C) LAMX, (D) RRFS1 QPF from 12 UTC June 29 to 12 UTC 03 July 2021.*

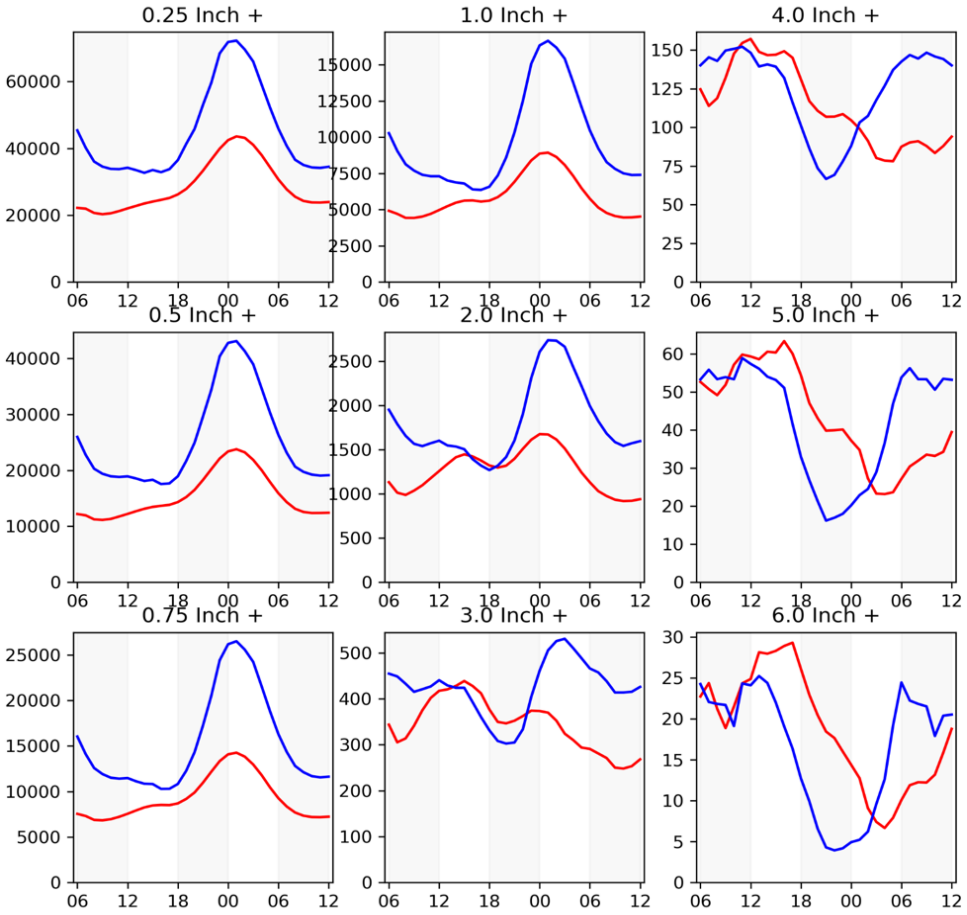


Figure 58: *Mean grid point counts of 6h MRMS QPE (blue) and 00z HRRR QPF (red) from 10 June to 31 July 2021. Thresholds (also listed above each graph) from top to bottom and left to right are as follows: 0.25", 0.5", 0.75", 1", 2", 3", 4", 5" and 6".*

The daily median of the max 6 hour QPF compared to the median max 6h QPE from MRMS (Fig. 61) also supports that the FV3-CAMs are more likely to forecast higher magnitude rainfall. The max 6h median for MRMS (Fig. 61A) is roughly 5.5 inches and the HRRR QPF (Fig. 61B) is similar, hovering around 6 inches out to f36. However, the LAMX (Fig. 61C) median is around 7.5 inches, increasing to ~8.5 inches at 00 UTC before dropping closer to 6 inches after 06 UTC. The RRFS1 (Fig. 61D) median max 6h QPF ranges from 10 to 15 inches and shows the most variability associated with forecast hour. Looking at the maximum of the maximums (black line), the HRRR has a comparable max to the RRFS1, nearing 35 inches in 6 hours, but the tendency of the maximum values across the CONUS differs greatly. The HRRR's maximum curve out to f36 has a similar shape to the MRMS while the RRFS1 maximum value with time does not resemble the MRMS pattern. In fact, it's maximum 6h QPF exceeds 18 inches at all forecast hours. Meanwhile, the LAMX has a maximum value closer to the observed maximum (25 inches to 21 inches respectively). However the pattern of the maximum curve does not resemble MRMS, staying relatively flat out to 03 UTC. This apparent inability of the RRFS1 and LAMX to forecast the diurnal variation in rainfall is troublesome as well.

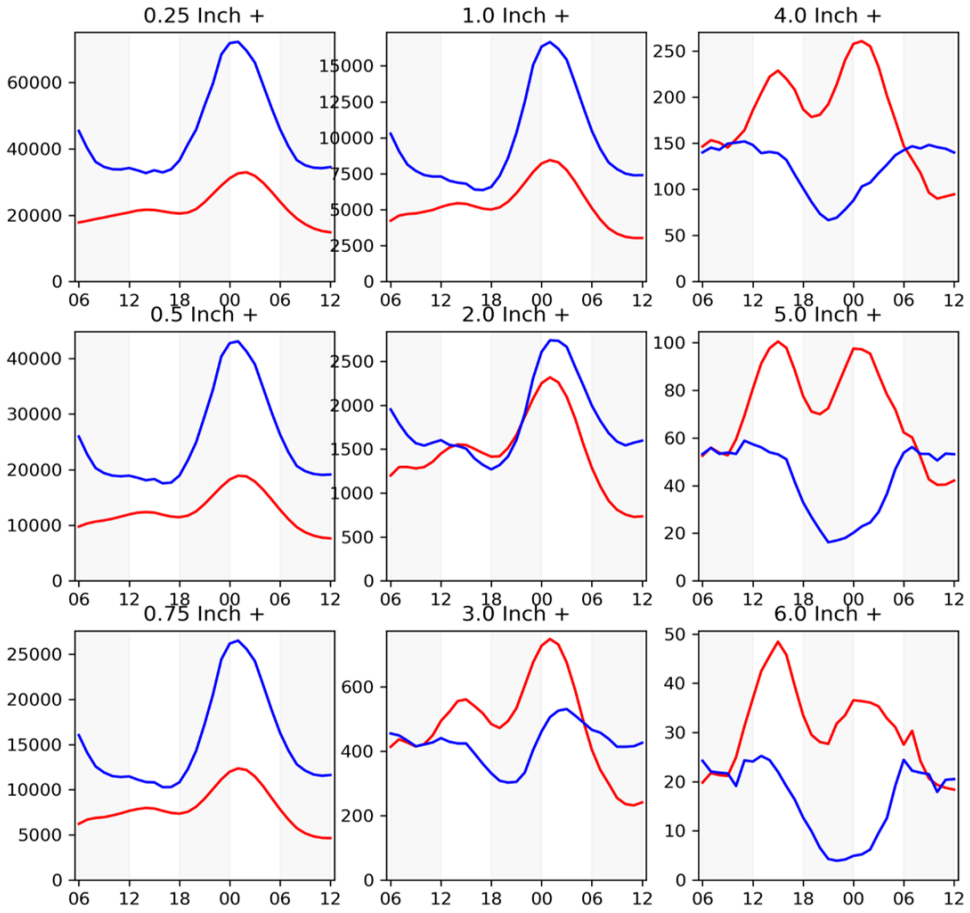


Figure 59: *Same as Fig. 58 but for the 00z LAMX (red).*

Finally, the PDs (Figs. 54 and 55) also show that the LAM and LAMX forecasts are comparable to one another, especially when analyzing their 00z performance. At 12z, when looking across the span of FFaIR, there appears to be more of a difference between the two models. However, when the analysis time period expands to two months (Fig. 56), the LAM and LAMX performance are similar. Again, the similarities between the two models is encouraging as the only difference between the two is the domain size. Focusing on the critical success index (CSI) for the half inch threshold at 00z, the RRFS1 was the worst performer[38] during FFaIR and the HRRR out performed all the models. At 12z, the RRFS1's CSI increased from 0.2 to 0.26 and is comparable with the LAM. LAMX and NAMnest. The LAMDAX was the worst performer at 12z during FFaIR but when looking over the whole two months of June and July it becomes comparable to the other LAMs. This large difference in performance is likely because the 12z LAMDAX data was missing 12 of the 20 days of FFaIR. Lastly, although all models evaluated tended to have an increase in CSI from the 00z to 12z cycle, none were as large as the RRFS1, which confirms the participant's observations that the RRFS1 12z run was nearly always significantly improved over its 00z run.

[38] This is true also for the whole June thru July time period.

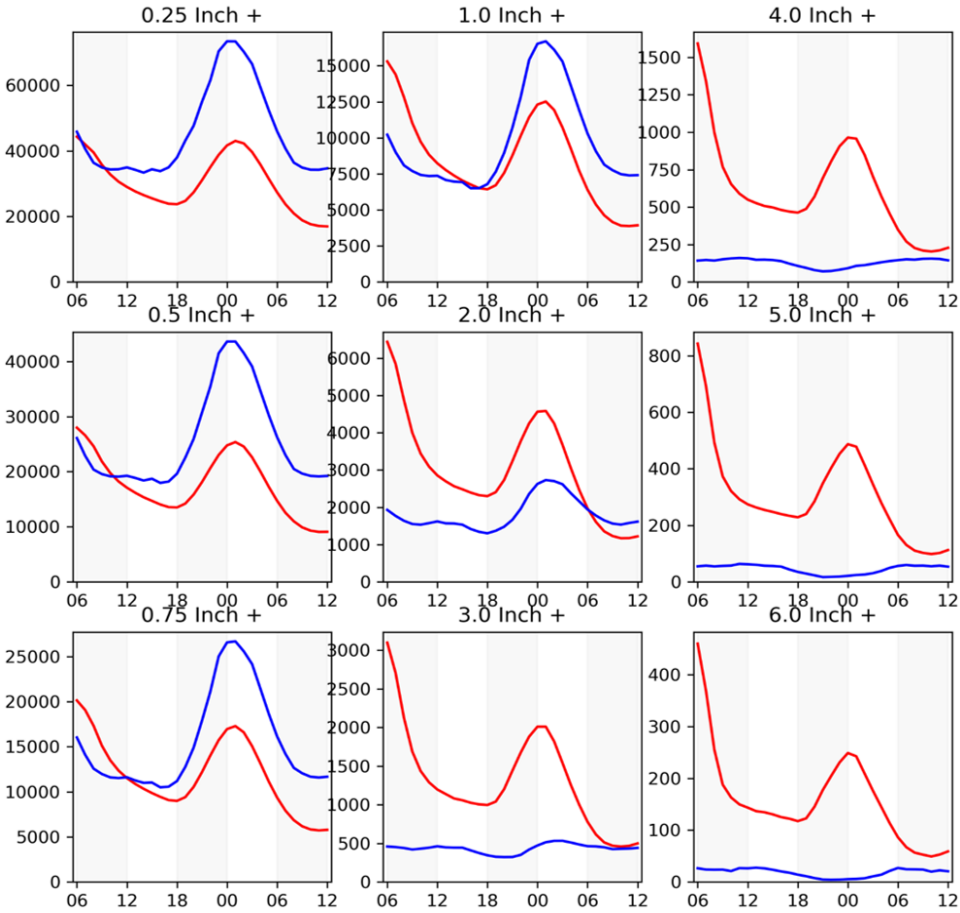


Figure 60: *Same as Fig. 58 but for the 00z RRFS1 (red).*

4.1.1.1 Quick Summary

It is apparent based on both the subjective and objective evaluation of the FV3-CAMs that the wet bias noted last year in FFaIR is still present. This is especially true when looking at precipitation totals exceeding 2 inches in both 6 and 24 hours. However, all the models, including the operational models, had a tendency to under forecast precipitation at thresholds less than 1 inch. Additionally, like last year, there continues to be an overdevelopment of popcorn convection in the FV3-CAMs. However, participants did note they felt that for the EMC models (LAM, LAMX, and LAMDAX) the QPF totals within these cells, as well as the overall wet bias, was less extreme than the EMC FV3-CAMs evaluated last year. When looking at the PDs for the FV3-CAMs last year (not shown[39]) compared to this year, overall there was a slight decrease in the wet bias in the FV3-CAMs at all precipitation thresholds. The SARX from last year has the same configuration as the LAM this year, with the exception of the ICs/LBCs being from the GFSv15 rather than GFSv16 and an increase in the vertical resolution from 60 to 65 levels. For FFaIR only dates, at the 1 inch threshold, the SARX had a bias around 1.5 and CSI around 0.13

[39] Can be seen in Fig. 31 of the 2020 FFaIR Final Report (https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf).

whereas the LAM had a bias around 1.2 and CSI of 0.15; a similar trend is seen when looking at the extended period between June and July. It is difficult to say if the change in vertical resolution and the updated GFS are the cause of the lower bias as it is possible changes not related to microphysics, boundary conditions, and vertical levels occurred that the FFaIR team is unaware of. Additionally, although the error introduced in the LBCs appears to have overall had only a small impact on QPF in Fig. 46, since the change in the wet bias was relatively small it is possible the error helped drive that change. Furthermore, the general weather patterns between the four weeks of FFaIR last year were noticeably different, making it difficult to make a reliable comparison between 2020 FV3-CAMs and the 2021 FV3-CAMs since last year had many MCS/MCV events, while this year there were very few of these and the Southwestern Monsoon dominated half FFaIR. Either way, the wet bias and aggressive forecasting of popcorn convection from the FV3-CAMs remains a concern of the FFaIR team and one that must be addressed before proceeding forward with implementation of the RRFS system.

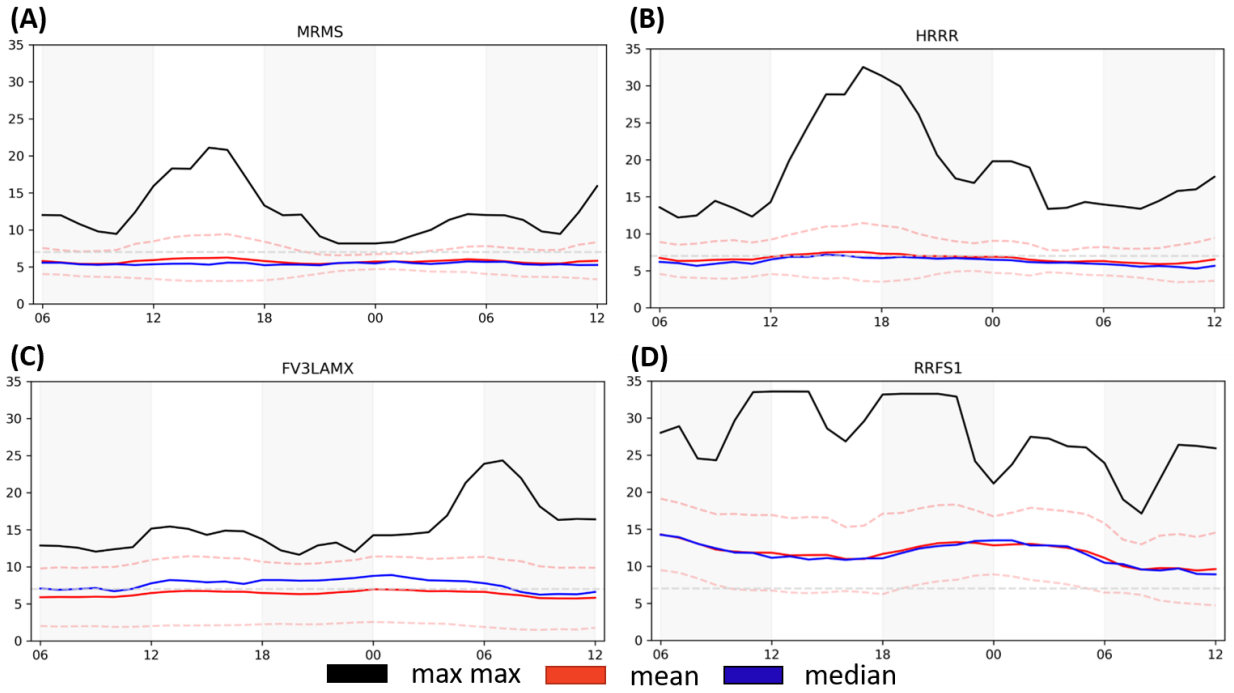


Figure 61: *Maximum 6h (A)MRMS QPE and (B) HRRR, (C) LAMX, and (D) RRFS1 QPF, averaged daily from 10 June to 31 July 2021. The black line indicates the maximum 6h QPF/QPE observed for each forecast hour over the course of the evaluation period (think of it as the max max). The dashed red lines are the standard deviation of the mean (red). The grey dashed line denotes the 7 inch threshold for comparison.*

4.1.2 Timing

The subjective analysis of product timing for QPF turned out to be more difficult for the participants to assess than the FFaIR team expected. This likely arose for many reasons, with the most predominant reason likely being a mix of the team's wording of the question and the question setup. To combat these shortcomings, an in-depth walk through on how to go about evaluating model timing was given. This included more thoroughly defining what exactly was meant by timing, which in this case meant the evolution of the event. For instance, if the focus of the day was QPF valid from 00-06 UTC, did that 6 h QPF pattern more closely match MRMS during that six hour window or did a six hour window before or after the valid time (ex. 02-08) more closely match MRMS valid 00-06 UTC? If the latter, could that 2 hour latency be traced back in time, suggesting a timing issue?

Additionally participants were instructed that if timing did not appear to be the culprit for a "poor" forecast then they did not need to check a box for one of the timing choices (refer to Figure 5 to see question setup and choices). In retrospect, an option for analysis choices should have been expanded to include things like "timing was not the issue" or "part timing/part convective evolution" or "timing issues differ across the area of interest". Choices like these might have helped address the feedback comments the participants gave at the end of the week like:

- "even if it was strongly forced... the northern part of the line is too fast while the southern part is too slow"
- "was hard to differentiate timing form positioning, because usually the positioning was off anyway"
- "The timing exercise was a bit tricky, especially when model errors were on exact placement and not timing."
- "The timing exercise was challenging during the week, due to deficiencies in the model forecasts. For example, it was hard to evaluate timing if the model QPF footprint was way off, or if precipitation was totally missing in the model"

Before discussing the results, it is important to note some additional caveats that likely impacted the results of this verification question. First and foremost, there was a large difference in the number of times each model was assigned to be evaluated due in part to both model availability, the number of participants each week and if models were assigned or chosen by participants[40]. For instance the HRRR was evaluated 104 times while the LAMDAX was only evaluated 26 times. In addition to this, there were times in which participants evaluated all of the

[40] Sometimes participants were instructed to choose their model to evaluate because the images for the model they were assigned were not yet available on the website. This was often due to the FFaIR team's machine's not having enough computing power to generate all the images needed for the website and for verification.

models rather than just one, even if a given model was not available that day. This was likely due to the gridded set up of the survey question; again refer to Fig. 5 to see the question's setup. Furthermore, on some days participants chose multiple choices for timing, i.e. they checked both the "too slow" and " too fast" boxes. The latter of these errors results in the percentages not equaling 100% when looking at the three timing choices: "timing is good", "too slow" "too fast". For instance, summing up the results for HRRR for timing is: good - 68.27%, slow - 25.96%, and fast - 15.38% gives a total of 109.61%.

The results from the subjective evaluation of model timing can be seen in Fig. 62. The HRRR was the most likely to have a QPF footprint and general pattern similar to observations. Additionally, when it came to timing the RRFS1 and HRRR forecasts were most likely to be considered to have good timing, at 69.84% and 68.27% respectively. The LAMDAX was the least likely of the four models to have good timing and 30.77% of the time the participants felt the timing of QPF was too slow. It was also the worst performer in correctly forecasting the location of QPF, with the participants feeling that the location was good less than 20% of the time.

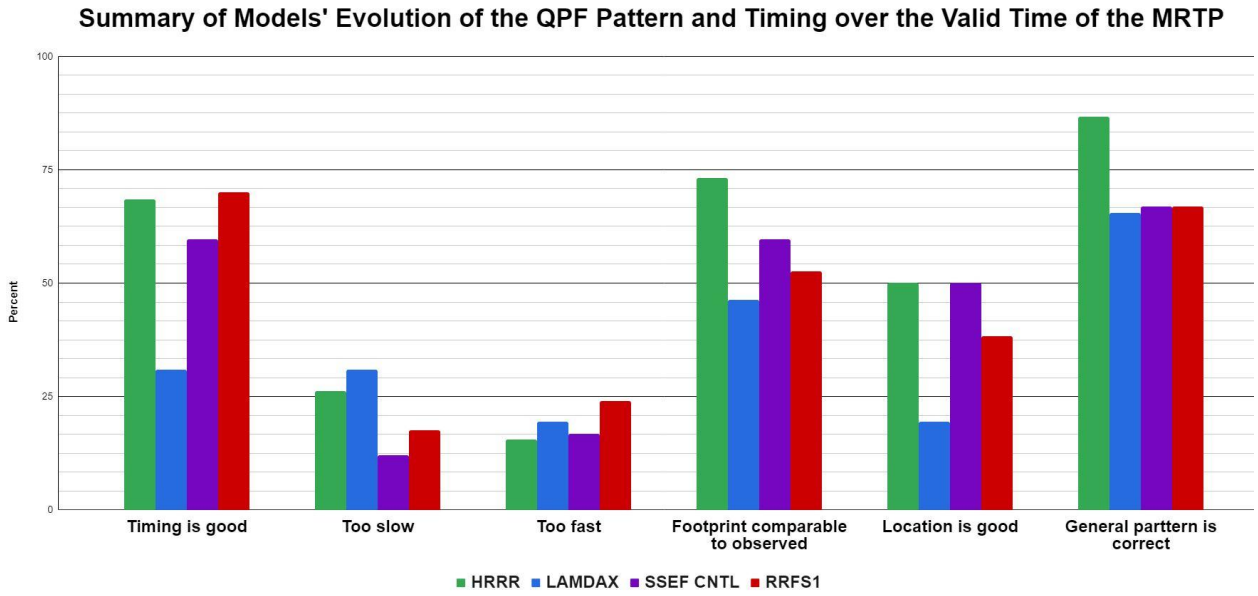


Figure 62: *Results from the subjective verification of model timing for the 00z HRRR (green), LAMDAX (blue), SSEF CNTL (purple) and RRFS1 (red).*

An example of a day in which a participant chose multiple timing options was for the 00z 24 June 2021 RRFS1 forecast valid 03-09 UTC 25 June 2021. Figures 63 and 64 show the 6 hour QPF verification image valid at the start (03 UTC) and the end of the forecast (09 UTC). Figure 63 is something the participant might look at to see how well the model was performing leading up to the period of interest while Fig. 64 is the image the participant would use for initial evaluation of the model's timing for the valid time period. For this day, a participant who chose two timing options ("timing is good" and "too fast") wrote:

> "The timing was a little quick, by 1-2 hours and while the pattern and footprint were similar, the overall extent was greater than what the model had initially and then the area was displaced to the east-southeast in the later runs (i.e. the area got worse with time). Generally the footprint was pretty good, it even seemed to pick up on the convection in KS."

Interestingly, a different participant evaluating the RRFS1 as well felt that the model timing was "too slow". That participant noted:

> "RRFS1 was a bit slower to develop the MCS in MO, but it seemed to eventually catch up to the obs. It also never developed the narrow band of heavy rainfall in KS. RRFS timing with the heavier precip out west seemed better"

These are just two responses for one event, an event that had a large footprint and strong forcings, nonetheless, still it was difficult to get a clear sense of timing. For reference, the responses from the other two participants that evaluated the RRFS1 for this event were one for "timing is good" and one for "too fast".

Despite the difficulties that arose from this question, when participants at the end of the week were asked about their general thoughts and comments for the timing question the response was mostly positive. They noted that although the exercise (question) was difficult they felt it was an interesting way to evaluate model performance and that analysis like it "could clue the developers to spin up, or progression issues with some of the experimental guidance." Multiple participants noted that it was a good addition to the verification sessions and that it allowed for an in-depth assessment of the 6 h QPF and to identify possible model biases. Therefore, even though the subjective verification for QPF timing was a mixed bag this year it has proved a good stepping block to further refine ways to evaluate model performance in a timing sense.

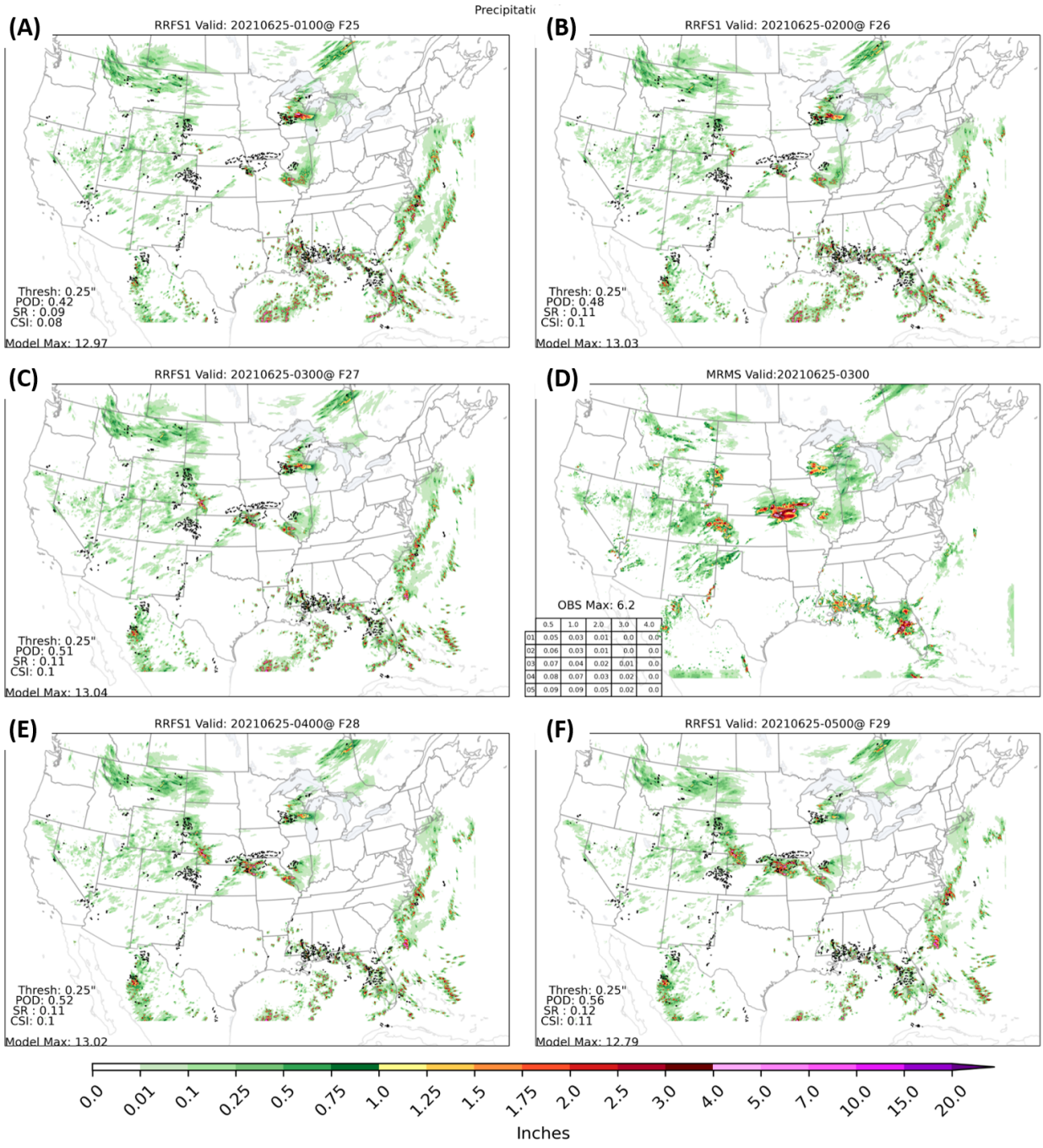


Figure 63: *The image for the subjective evaluation of the RRFS1 timing performance valid at 03 UTC 25 June 2021. Refer to Fig. 6 and the discussion surrounding it on how to interpret the image.*

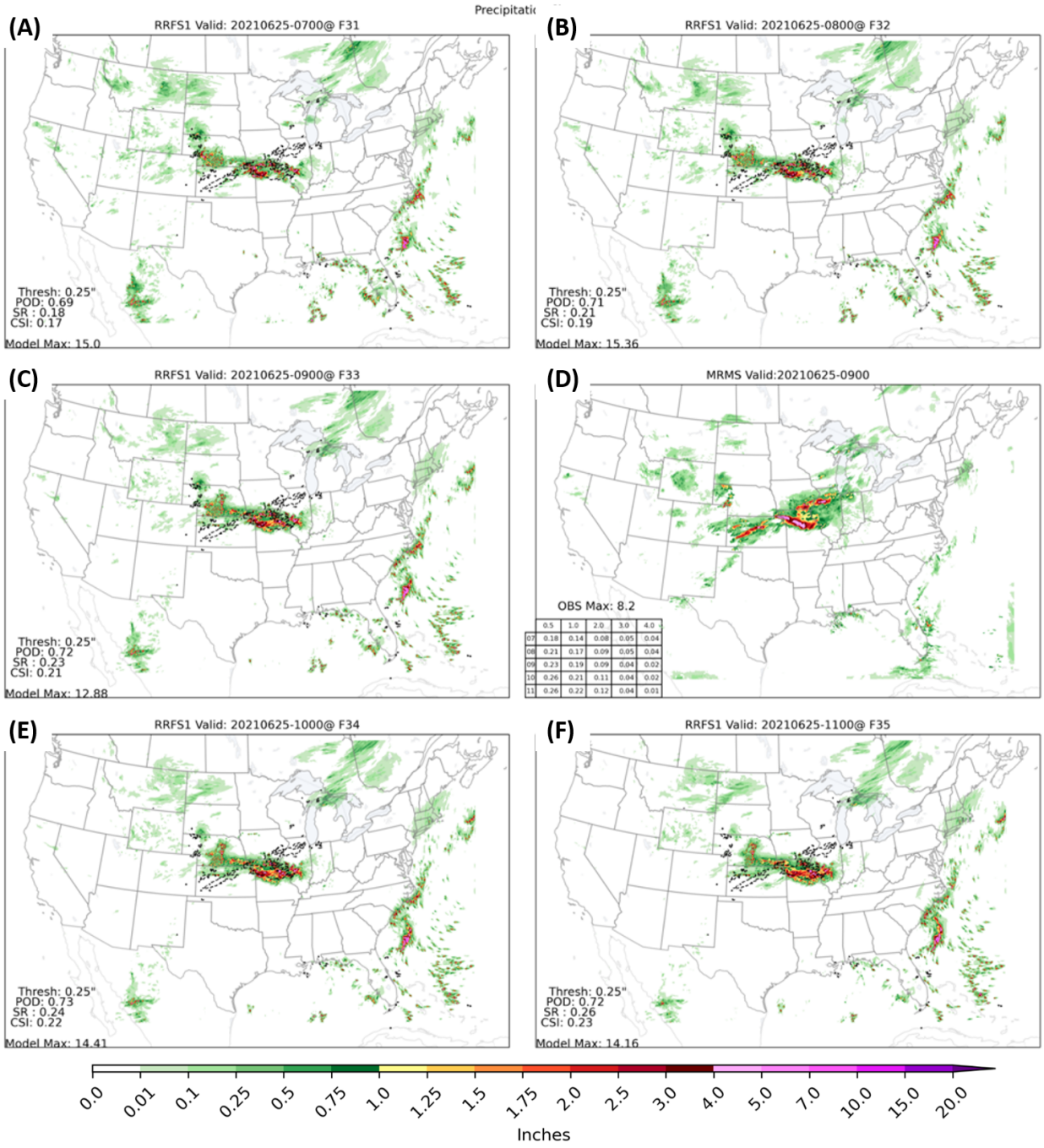


Figure 64: *Same as Fig 63 but valid at 09 UTC 25 June 2021.*

4.1.3 Instantaneous Precipitation Rate

Subjective verification of p-rate provided two distinct results, 1) the spatial extent of model p-rate is less than that of the MRMS and 2) the experimental models have higher, in some cases significantly higher, p-rate values compared to MRMS and the HRRR. The former of these two results largely drove how participants answered the evaluation question asking if the model p-rates was wetter, drier or about the same as MRMS p-rates. As can be seen in Fig. 65

participants overwhelmingly felt that the models were drier than observations, though the HRRR tended to be less dry than the experimental models. The participant comments for this analysis regularly discussed the lack of coverage leading them to say the model was dry despite high p-rate values. For instance, one participant noted "Rainfall rates under the specific cores seemed to over-estimate rainfall in the heavier cells, but the coverage leans me to rate as a dry bias." A zoomed in image of the p-rate verification valid for the day this comment was written can be seen in Fig. 66. Focusing on the RRFS1 p-rate (Fig. 66C), coverage of p-rates was noticeably lacking, especially from Saint Louis, MO to southern Michigan. However, across southern Missouri, p-rates are generally above 3 in/hr and the maximum p-rate (indicated by the gray box) was 32.46 in/hr. The maximum p-rate for MRMS on the other hand, was 7 in/hr from MRMS.

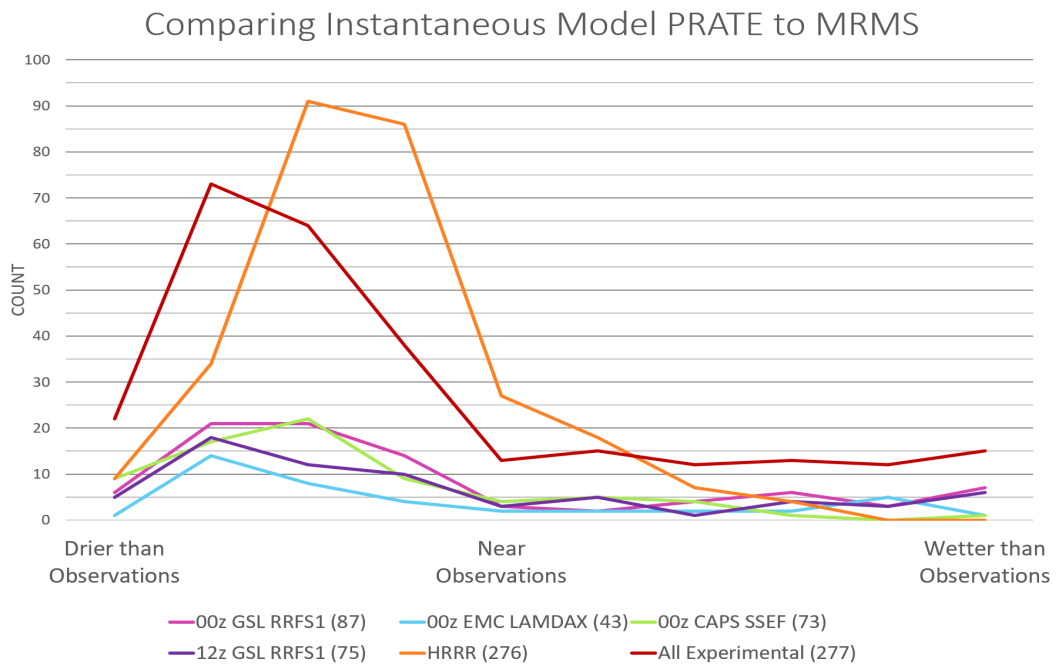


Figure 65: *Results from the subjective verification for p-rate for showing the number of times each model forecast was drier, near, or wetter than observations for the 00z RRFS1 (pink), 12z RRFS1 (purple), 00z LAMDAX (blue), 00z and 12z HRRR (orange), 00z SSEF CNTL (green) and the sum of all experimental models (red).*

This subjective verification question is another example of the importance of the question's wording and response choices. The hope when this verification question was created was to tease out how p-rate values themselves compared with MRMS and the impact of spatial coverage was not considered. Perhaps a better way to word the question to gather the information we were trying to get would have been to ask if the overall magnitude of the model p-rates were more/less intense than the observed p-rates. It also might have been helpful to the participants if a different colorbar and scale for plotting rain rate was used rather than the QPF colors and scale. However, p-rate analysis was new to FFaIR this year and the wide range of values, especially the number of values exceeding the upper limit (20 in/hr) of the colorbar, was not anticipated.

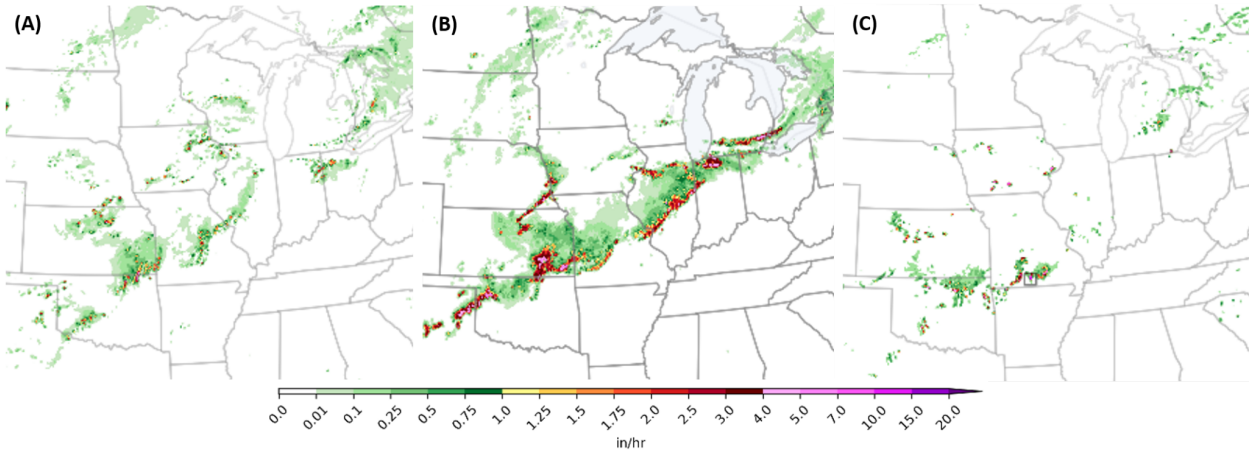


Figure 66: *P-rate from (A) 00z HRRR, (B) MRMS, and (C) 00z RRFS1 valid 05 UTC 29 June 2021. Grey box indicates the p-rate maximum for the CONUS.*

In regards to the second finding involving p-rate magnitude, a summary of model and MRMS maximum p-rates can be seen in Fig. 67. The HRRR average maximum p-rate is consistent with the maximum average from MRMS, hovering around 8 in/hr. The LAMs from EMC have a higher average maximum p-rate but are generally consistent with one another, roughly staying between 20-30 in/hr throughout the forecast. However, unlike the HRRR and MRMS the diurnal cycle can be seen in the maximum averages from the LAMs, peaking between 18z and 22z for both the 00z and 12z model initializations. A similar diurnal cycle can also be seen in the RRFS1 average maximum p-rate, with the average maximum p-rate ranging from 40-60 in/hr for the 00z model run and 55-35 in/hr for the 12z run. Additionally, for the 00z RRFS1 run, there are noticeably higher p-rates at model initialization compared to the rest of the model forecast hours. This could suggest shortcomings in the DA processes used for the RRFS1.

Plotted along with the maximum average p-rate in Fig. 67 is the maximum p-rate for each forecast hour that occurred during the 2 months of analysis (June - July). Not surprising the HRRR maximum p-rates are of similar magnitude to MRMS, though there is some discrepancy. The HRRR maximums rarely exceed 16 in/hr while in the MRMS there are instances in which the maximum p-rate exceeded 20 in/hr, approaching 25 in/hr. Therefore there is a slight underestimate of the highest p-rates from the HRRR. This clearly is the opposite of what is seen in the LAMs and RRFS1. Maximum p-rates from the LAMs generally stay between 40-60 in/hr, with some instanstances of rates approaching 100 in/hr during the convective maximum part of the day. The RRFS1 maximum values are even larger, especially in the first 30 hours of the 00z forecast and in the first 12 hours of the 12z run where maximum p-rate magnitudes exceed 100 in/hr. Moreover, there were instances over the two months of analysis that the maximum p-rate exceeded 150 in/hr, with an occurrence of 300 in/hr.

The high bias seen in the LAMs and RRFS1 p-rates is troubling and seemingly unphysical. During FFaIR, there was discussion on what physically could be reasonable from

model output at such small timesteps. Afterall, as stated in Section 2.3.1, comparing model p-rates to MRMS is not a true apple to apple comparison, so it is possible that higher rainrates occur. However, other research has examined model p-rates with observational data that was not MRMS. For instance, Bao and Sherwood 2018[41] plotted tropical rain rates from Manus using an 1 min optical rain gauge showing rates up 12"/hr suggesting that the MRMS data is a good representation of the range of possible p-rates. Additionally, in their simulations using the WRF model, similar p-rate magnitudes to HRRR were found. Therefore, it is highly likely that the HRRR and the MRMS provides a good baseline for reasonable p-rates, even at small timesteps. The FFaIR team suspects there is a deficiency in the FV3 as it relates to instantaneous precipitation rates, and thus QPF itself, that needs to be identified and assessed before RRFS implementation can occur.

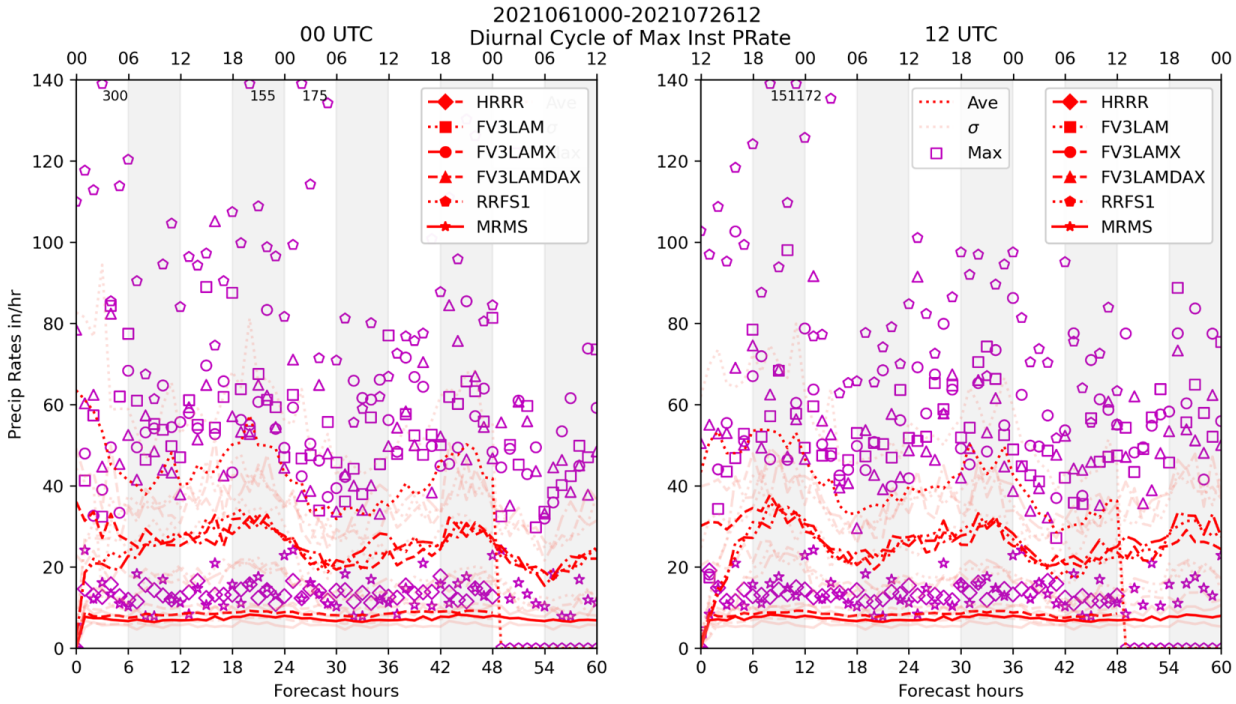


Figure 67: *Analysis of domain maximum p-rate from 10 June to 31 July 2021 for (left) 00z cycle and (right) 12z cycle. Dark red lines indicate each model/MRMS average maximum p-rate. Light red lines are each model/MRMS standard deviation of the maximum p-prate. The purple shapes are each model/MRMS maximum p-rate for each forecast time over the course of the evaluation period (think of it as the max max). Refer to the table for each model's/MRMS respective shape and line format.*

4.2 Ensemble Guidance

As stated in Section 2.3.1, this year participants evaluated the performance of ensembles through probability of exceedance thresholds, specifically exceeding 1in/3h and 2in/6h, for both the Day 1 and Day 2 forecast. The hope was to better understand how forecasters use probabilities and how they determine if probabilities are too high or too low. This proved to be a

[41] https://doi.org/10.1029/2018MS001503

more difficult task than the FFaIR team anticipated, likely due to the setup of the question. Although participants were asked if they felt the probabilities were high, low, or about right, information about the location and spatial extent of the probabilities was not asked. This shortcoming was apparent when listening to the participants during the verification discussion and when reading through their written comments. For example, comments like "RRFSCE has a 3-hour period where the 1-inch probabilities looked appropriate in magnitude, but displaced to the south of observations" and "(p)lacement was more egregious than percentiles, maybe. All models had high confidence in places with little precip and missed other heavier locations" were common throughout the experiment. Participants noted that displacement errors made it difficult to truly analyze whether or not the probabilities were representative of the rainfall risk.

Figure 68 shows the results of the ensemble analysis question for the Day 1 forecast for the HREF, RRFSCE, and SSEF and the Day 2 forecast for the RRFSCE and SSEF. As has been the trend in the past, the HREF was considered to be the best performing ensemble according to the participants. 57% (49%) of the time the participants felt that the HREF 1in/3in (2in/6h) probabilities were "about right" while the RRFSCE was felt to be "about right" 32% (26%) of the time. When comparing the exceedance performance of the 1in/3h to the 2in/6h, for both the HREF and RRFSCE, perception of probabilities was inconsistent between the two exceedance thresholds. The participants often felt that the 2in/6h probabilities were "low" (33% and 34%, respecitively) rather than "high" (10% and 17%, respectively). But the same sentiment was not felt for the 1in/3h probabilities. Instead the participants were more likely to select that the HREF 1in/3h probabilities were "high" 25% of the time. The RRFSCE 1in/3h probabilities were also more likely to be chosen as "high" than "low", but this was only by a slim margin, 25% to 22%. One possibility for this inconsistency was that the 3h QPE was not provided for comparison, only the 6h was. Though in some instances, participants did explain the reason for these differences. For instance, one participant wrote " HREF 1in3hr product seemed to be near right and provided useful information highlighting the areas of most concern, however that was pretty much all washed out in the 2in6hr product." The corresponding rates for the HREF were "about right" for the 1in/3h but "low" for the 2in/6h.

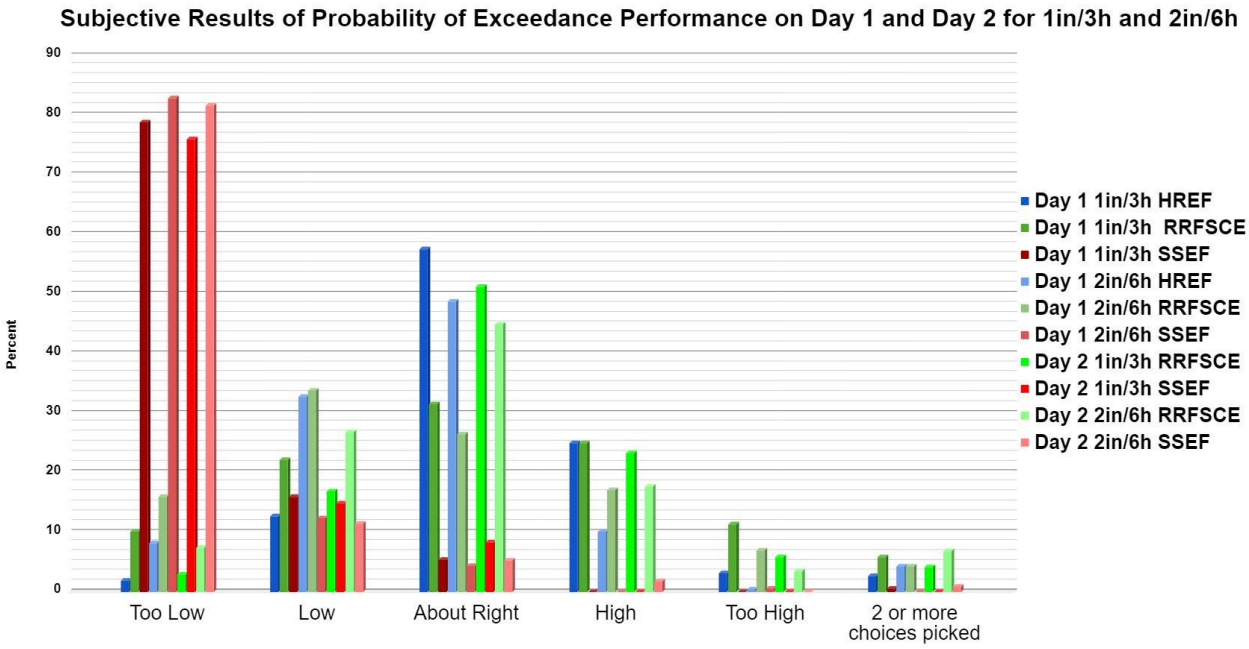


Figure 68: *Results from the subjective verification for the ensemble probabilities 1in/3hr and 2in/6h from the HREF (blue shades), RRFSCE (green shades), and SSEF (red shades) for the percent of the time a given high/low choice was picked.*

Figures 69-71 show the Day 1 ensemble verification images. Following the trend throughout FFaIR of the participants' having different perceptions on how the HREF and RRFSCE 1in/3h probabilities did compared to the 2in/6h probabilities, this day saw notable differences between the two probabilities. This was especially true for the HREF, where 13/15 of the participants felt the 1in/3h probabilities were "about right" which dropped to 5/15 for the 2in/6h probabilities. Meanwhile, for the RRFSCE 12/15 participants felt the 1in/3h probabilities were "high" or "too high" while the remaining 3 felt they were "low" or "too low". But for the 2in/6h probabilities these changed to 9/15 and 6/15 respectively. One participant explained this inconsistency in opinion, stating: "HREF 2 in 6 hr probs were too low in Ohio Valley given many areas received 2-4 inches. 1 hr in 3 inches probs of 50-70 were ok, maybe could be higher given obs of 2-4 inches."

Focusing on the SSEF's probabilities, over 75% of the time the Day 1 probabilities were "too low". These results however might be somewhat skewed for a couple of reasons. The greatest reason being that it wasn't until July 15 that the probabilities were calculated using the same methodology as the other two ensembles, with point probabilities rather than neighborhood probabilities being provided during the first part of Week 2. Additionally, the SSEF was not available to evaluate all of Week 1 and therefore was evaluated less often than the HREF and RRFSCE. SSEF was scored 188 times, compared to the 261 (HREF) and 257 (RRSFCE) times the other two ensembles were scored. When looking just at the days when the SSEF probabilities

were calculated the same way as the other two ensembles[42], the overwhelming majority of the participants felt that both the 1in/3h and 2in/6h probabilities were "too low" (74% and 84% respectively). For the same time period the HREF and RRFSCE were felt to be "too low" 4% and 10% of the time for 1in/3h and 14% and 17% of the time for 2in/6h. Therefore, although the aforementioned issues impacted the results of the SSEF analysis, the model probabilities overall are too low to be of use according to the participants. The valid date in Figs. 69-71 is once the SSEF probability calculation method was the same as the HREF and RRFSCE. Comparing their 2in/6h probabilities, one can see that both the probability magnitude and coverage are noticeably lower from the SSEF compared to the other ensembles.

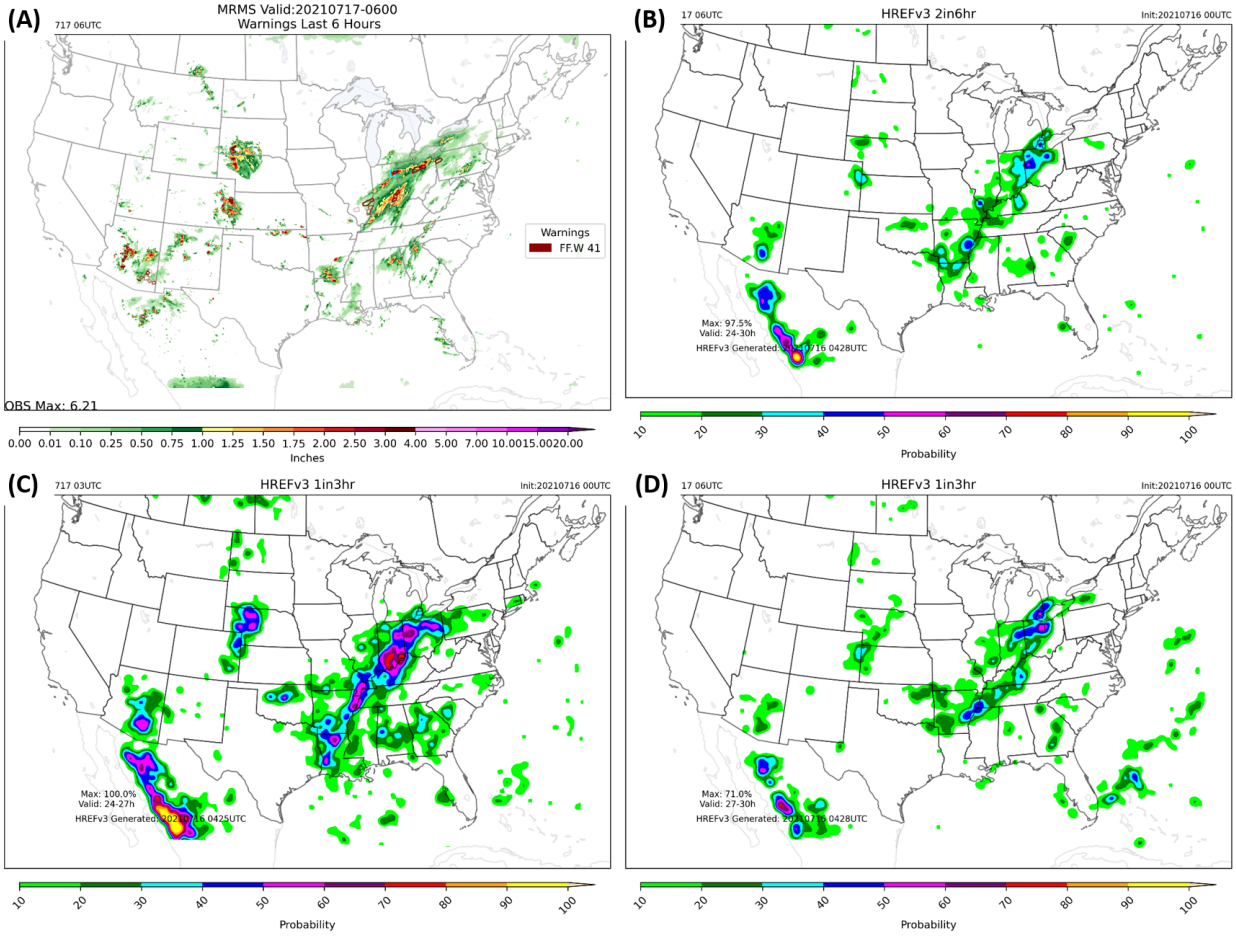


Figure 69: *(A) 6h MRMS and (B) HREF 2in/6h probabilities valid 00 UTC to 06 UTC 17 July 2021. HREF 1in/3h probabilities valid (C) 00 UTC to 03 UTC and (D) 03 UTC to 06 UTC 17 July 2021.*

The PDs for the ensemble mean, probability matched mean (pmm), and local probability matched mean (lpmm) 24h QPF at 0.5, 1, 2, and 3 inches can be seen in Fig. 72. Using CSI to define performance, each HREF mean outperforms the respective means of the RRFSCE and SSEF at all thresholds. For instance, at the 1 inch threshold the HREF pmm is 0.24 while the

[42] During this time period, the number of times scored were: HREF - 92, RRFSCE - 91, and SSEF - 88.

SSEF pmm is 0.21 and the RRFSCE pmm is 0.17. Additionally, aside from the 3 inches threshold, the SSEF means outperform the RRFSCE means. The poor performance of the RRFSCE compared to the other two ensembles is likely driven by the RRFSCE being under dispersive.

The quantitative results differ greatly from the subjective verification results, in which the participants felt the SSEF severely underperform, with respect to QPF probability of exceedance thresholds, compared to the HREF and RRFSCE. Even more, the participants often noted the RRFSCE seemed comparable to the HREF. Obviously, the mean products and probability of exceedance are not directly comparable, however it is interesting that the general participant feeling of the RRFSCE was positive in subjective verification but contingency table verification metrics suggest the RRFSCE performed quite poorly during FFaIR. This discrepancy likely arises from the low spread in the RRFSCE, which leads to higher probabilities. The higher probabilities look comparable to the HREF when the RRFSCE has a good handle on the event, thus the participants favorable perception of it.

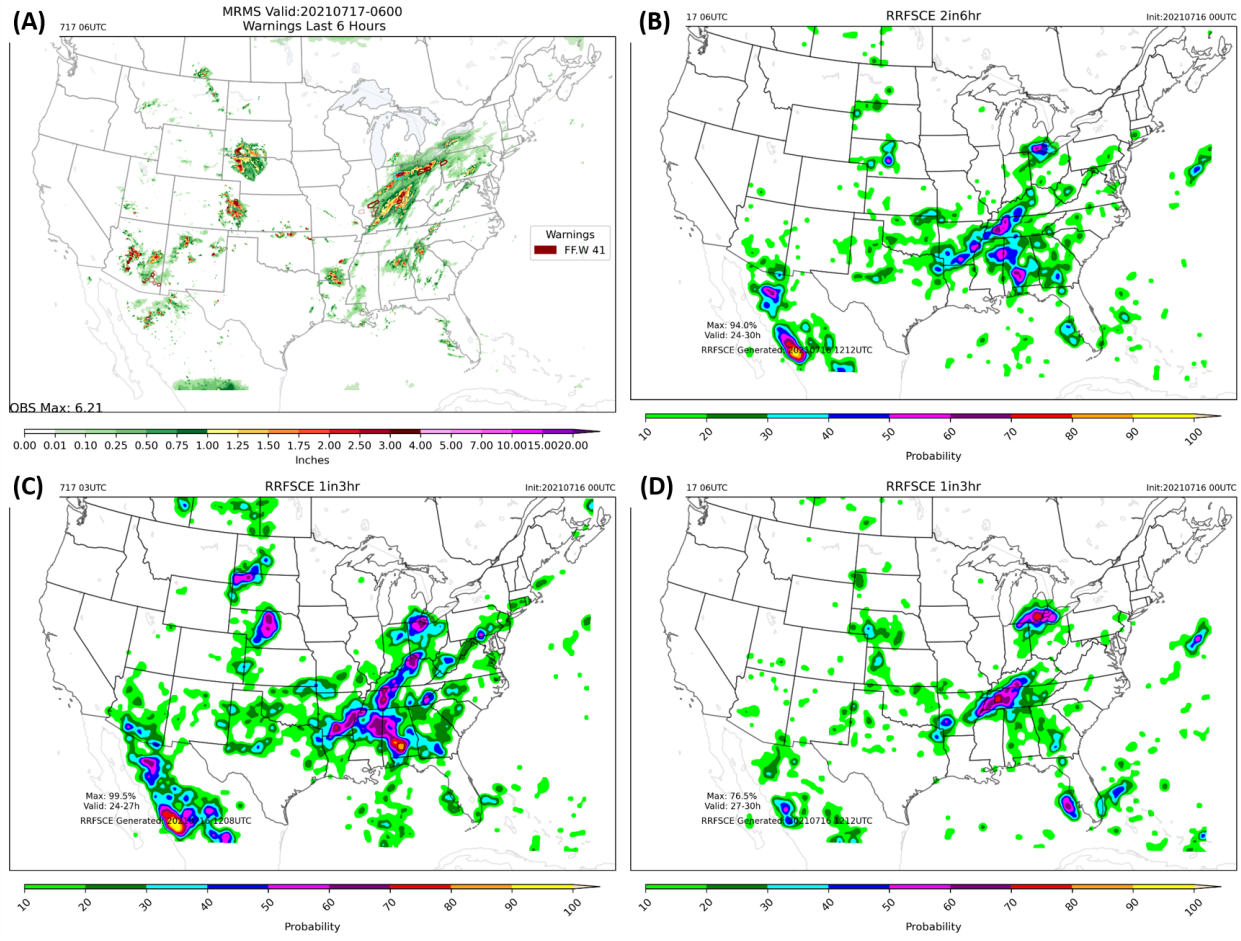


Figure 70: *Same as Fig. 69 but for RRFSCE.*

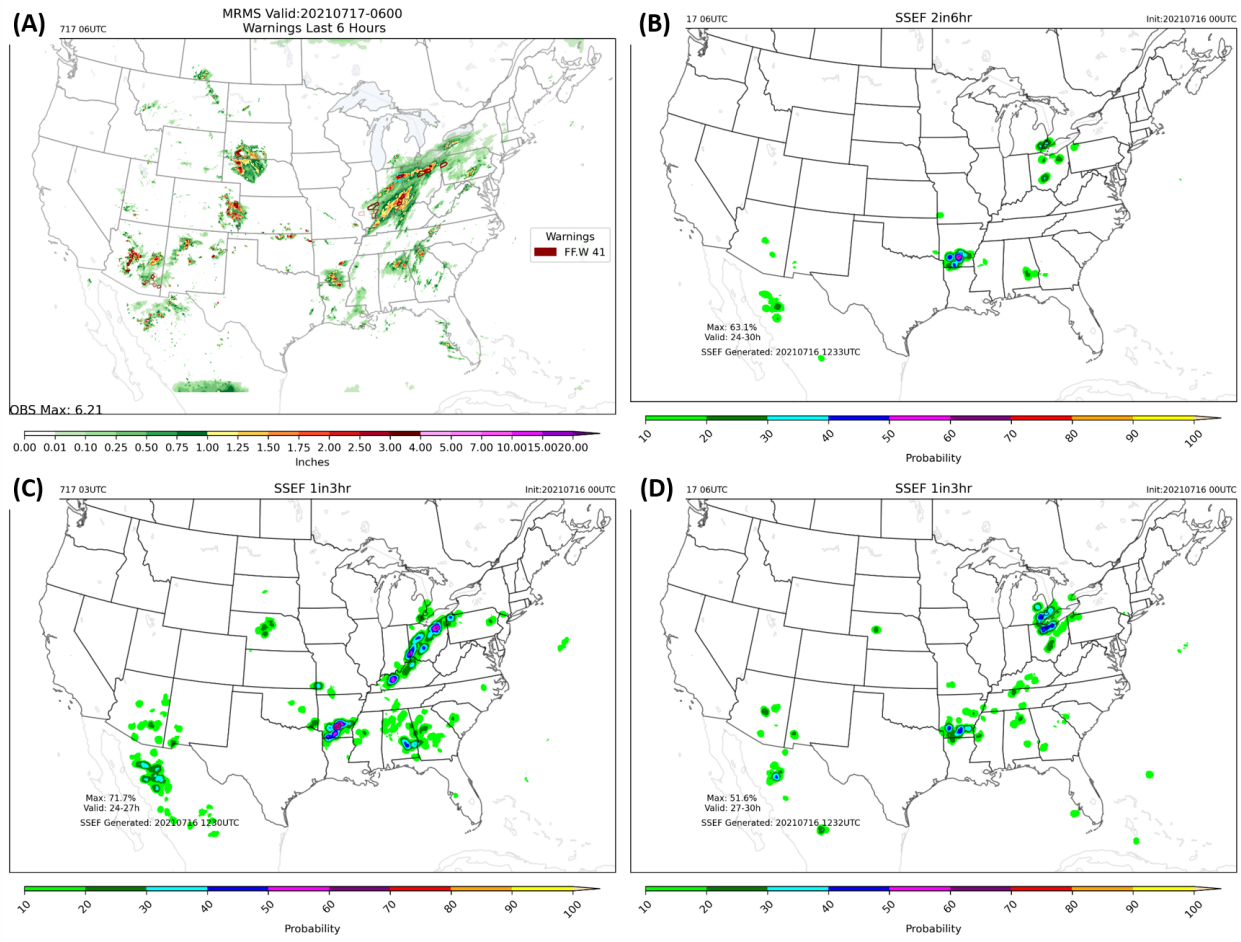


Figure 71: *Same as Fig. 69 but for RRFSCE.*

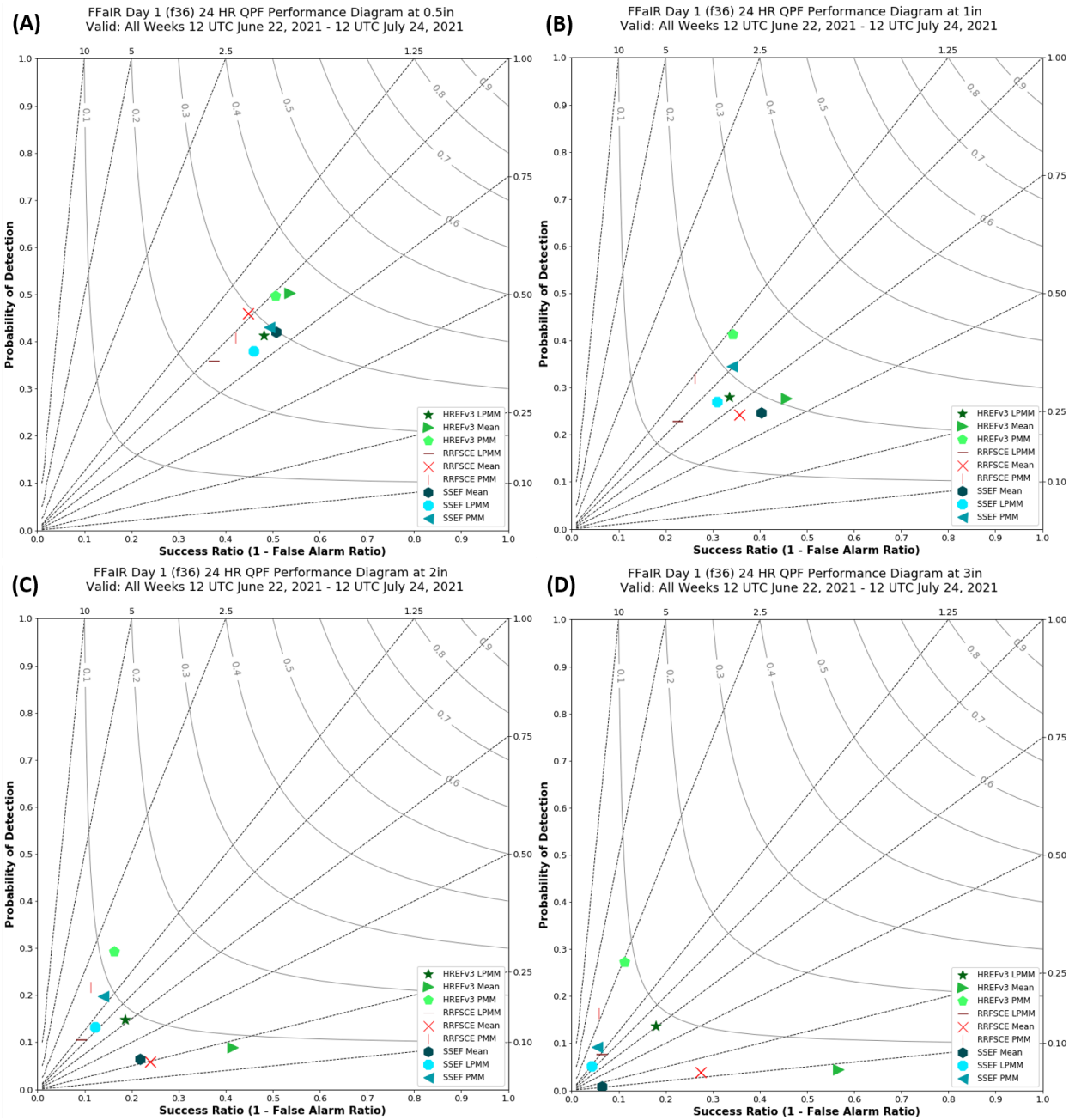


Figure 72: *Performance diagrams for Day 1 valid for the FFaIR days, from June 22 to July 23, 2021 for the ensemble mean, PMM, and LPMM. Precipitation thresholds are for: (A) 0.5 inches, (B) 1 inch, (C) 2 inches and (D) 3 inches. Ensemble means from each ensemble are the same color shades, i.e. all HREF means are green, RRFSCE are red and SSEF are blue.*

4.3 All Things ERO

For FFaIR this year, CSU provided six different configurations of their "First Guess" Day 1 ERO products (hereafter CSU EROs), 4 versions trained on the NSSL model, one trained on the HRRR and a Blended version that is a blend of the NSSL2, HRRR, and GEFS[43] CSU EROs. Explanation of the differences between the NSSL versions can be found in the FFaIR Operations Plan while Table 4 summarizes these differences. The CSU EROs are valid from 12 UTC to 12 UTC and are based on the 00z runs of each of the models. This differs from the Day 1 WPC ERO, FFaIR ERO, and ARI-ERO which are valid 16 UTC to 12 UTC.

Table 4: *Classification of the CSU First Guess ERO NSSL versions.*

FFaIR Naming Convention	Detailed Name	Description
NSSL2	NSSL-sptavg	Spatially averages predictor information.
NSSL3	NSSL-sptavg-landsea-mask	Spatially averages predictor information and masking applied in regions that border water.
NSSL4	NSSL-sptavg-landsea-mask-params	**RF parameters differ slightly from all other NSSL versions**. Spatially averages predictor information and masking applied in regions that border water.
NSSL5	NSSL-tempavg	Temporal dimension is averaged.

4.3.1 CSU "First Guess" EROs and FFaIR ERO

A summary of the subjective scores can be found in Fig. 73. Comparing the FFaIR ERO to the CSU EROs, the FFaIR ERO was subjectively the best. It never received a score of 1 or 2 (considered extremely poor forecast) and the overwhelming majority of the scores were a 7 or higher (77%), with roughly 10% of the scores being a 9 or 10 (extremely good forecast). The NSSL2 was considered to be the second best performer overall and therefore the best subjectively among the CSU EROs. NSSL3 and NSSL4 subjectively performed similar to one another, though it appears that the participants felt that the NSSL3 generally was more "average" than the NSSL4, which was more likely to score both above 6 and below 5 more often than the NSSL3. For instance, just over 50% of the time the NSSL3 received a score of 5 or 6 compared to 44% for the NSSL4. But the NSSL4 was scored 7 or higher 36% of the time compared to the NSSL3's 32% and 20% of the time was scored as a 4 or lower compared to 18% for the NSSL4. NSSL5 was the worst performer of the NSSLs while the HRRR was considered the worst performer overall.

[43] The GEFS ERO was not evaluated in FFaIR this year, but has been in the past and is used operationally by the WPC forecasters.

The BLEND's distribution of the subjective scores provide more detail, combined with participant's comments, about it's performance than the scores themselves do. Notice in the top image of Fig. 73 that the shape of the score distribution is nearly symmetrical, suggesting that participants felt the BLEND was nearly as likely to perform well as it was to perform poorly. This result can actually be traced back to the change in the focus of the heavy rainfall from mostly Eastern CONUS events for the first two weeks of FFaIR to largely Monsoon driven (Western CONUS) events the last two weeks of FFaIR; Figs. 23 and 24 show that at least a slight risk was issued in the FFaIR ERO over the last two weeks of FFaIR. The HRRR ERO, which is a component of the BLEND, had a dry bias in the southwestern US. This was noted by the participants when looking at both HRRR QPF guidance and the HRRR ERO. The dry bias from the model QPF over this region impacted the HRRR ERO forecasts which in turn impacted the BLEND ERO. The impact of the Monsoon on both the HRRR and especially the BLEND subjective scores can be seen in Fig. 74. Just focusing on the average scores, during the first two weeks of FFaIR, the BLEND had the highest average score, but during the second two weeks, it fell to being one of the worst forecasts. This change in the performance between the two halves of FFaIR could also be seen in the stark difference in how participants talked about the BLEND forecast during the first half of FFaIR compared to the second half. Comments like "The CSU-first guess Blend ERO is definitely the high performer of this group" and "The blend was the best again it captured most of the area of concern" were abundant during the first two weeks. However, comments like "Blend did bad because of the HRRR" and "HRRR and Blend missed everything in the SW" dominated the discussion the second two weeks, especially Week 4.

Figure 75 shows an example of how the HRRR ERO dry bias across the southwest impacted the BLEND forecast; the 24 h MRMS and HRRR QPF can be seen in Fig. 76 to show the dry bias in the HRRR QPF across the southwest. As can be seen the HRRR ERO has only a small area of Marginal risk across AZ, while the GEFS and NSSL2 EROs have a slight risk across much of AZ and into NM. However, the lack of risk from the HRRR product damps the slight risk greatly and results in only a small area of slight in AZ. Practically Perfect suggests a Moderate Risk was valid for this day, so all the guidance that went into the BLEND was low but the HRRR had the worst forecast across the region. On this day the weights that created the BLEND were: GEFS - 0.3457, NSSL2 - 0.3297, and HRRR - 0.3246. Since the HRRR model itself appears to have a consistent dry bias over the southwest during the monsoons season, it might be beneficial for the CSU team to reevaluate how the daily weights are determined, perhaps dividing the CONUS into an east and west region and setting the weights based on each model's running verification regionally instead CONUS wide.

Looking at performance metrics, the FFaIR ERO compared to the Operational ERO performed relatively similar to one another. In Fig. 77, it can be seen that the FFaIR ERO performs slightly better based on the Bulk Brier Score (BS) but slightly worse when using Bulk Area Under the Curve Receiver Operating Characteristics (AUROC). The difference in outcomes between the two metrics could be contributed to by the expertise the WPC forecasters have for

the monsoon over the FFaIR forecasters. Looking at the FFaIR ERO Brier Skill Score (BSS), Fig. 78A, it can be seen that the FFaIR ERO generally had around the same to slightly better skill than the Operational ERO during the first three weeks, but its skill fell below the Operational ERO during Week 4, when the Monsoon ramped up. The reliability, in the form of fractional coverage, of the Operational and FFaIR EROs can be seen in Fig. 78B. Overall, the Marginal and Slight Risks were relatively comparable to one another. Meanwhile the FFaIR ERO appears to be slightly more collaborated than the Operational ERO for the Moderate Risk, which is likely driven by the fact that the FFaIR Moderate Risks were generally larger (35.6% larger) than the Operational EROs and were drawn more frequently. This could suggest the Operational ERO Moderate Risks drawn during FFaIR are either not large enough or were not issued enough; this is consistent with the results from Erickson et al. (2021). Over the course of FFaIR, participants only issued one High Risk, on July 23, 2021, while no High Risk was issued by the WPC. Figure 79 compares the two EROs along with the Practically Perfect verification and MRMS for the day. As can be seen, the High Risk did not verify.

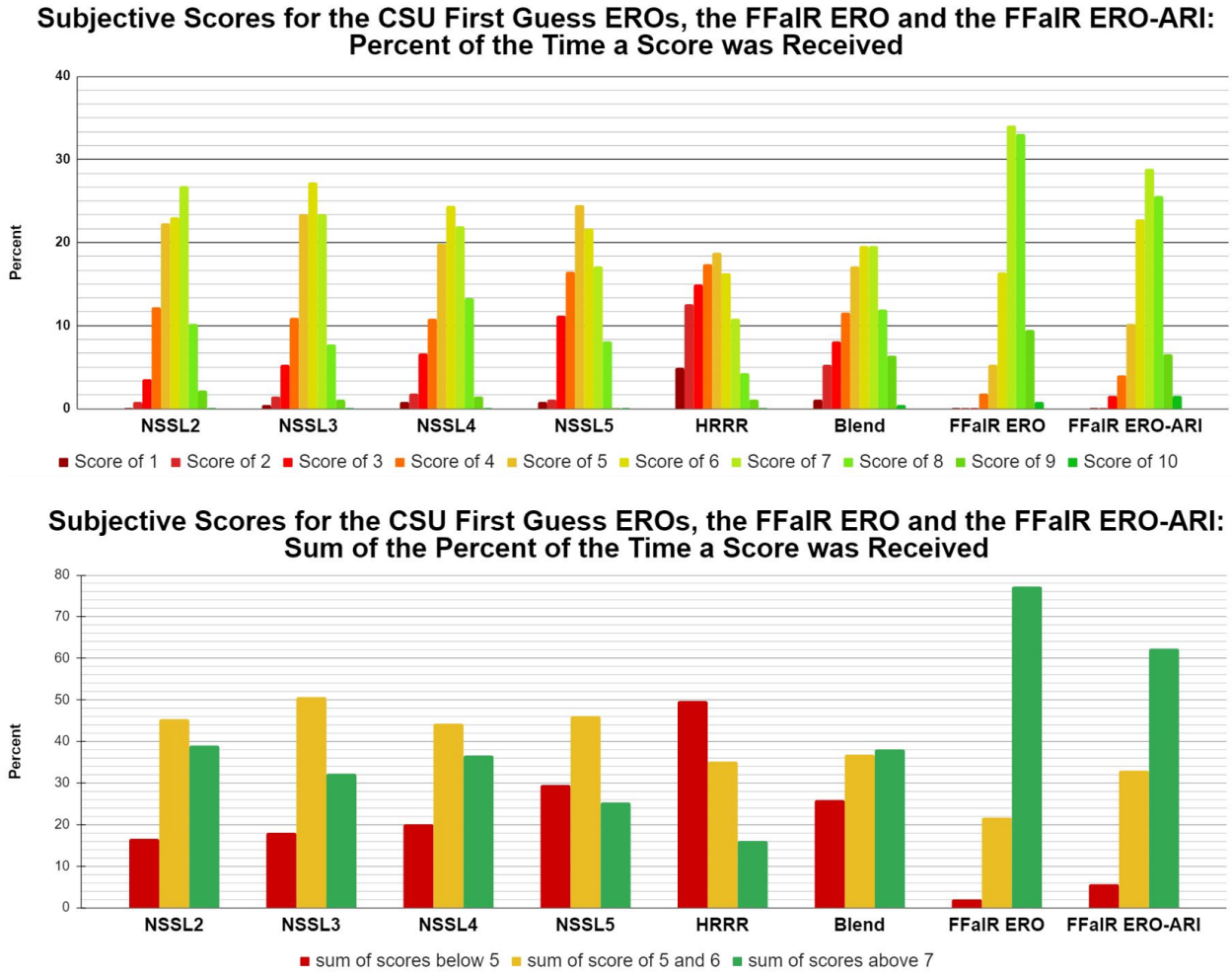


Figure 73: *Same as Fig. 47 but for the subjective analysis of the CSU EROs, FFaIR ERO and FFaIR ARI-ERO.*

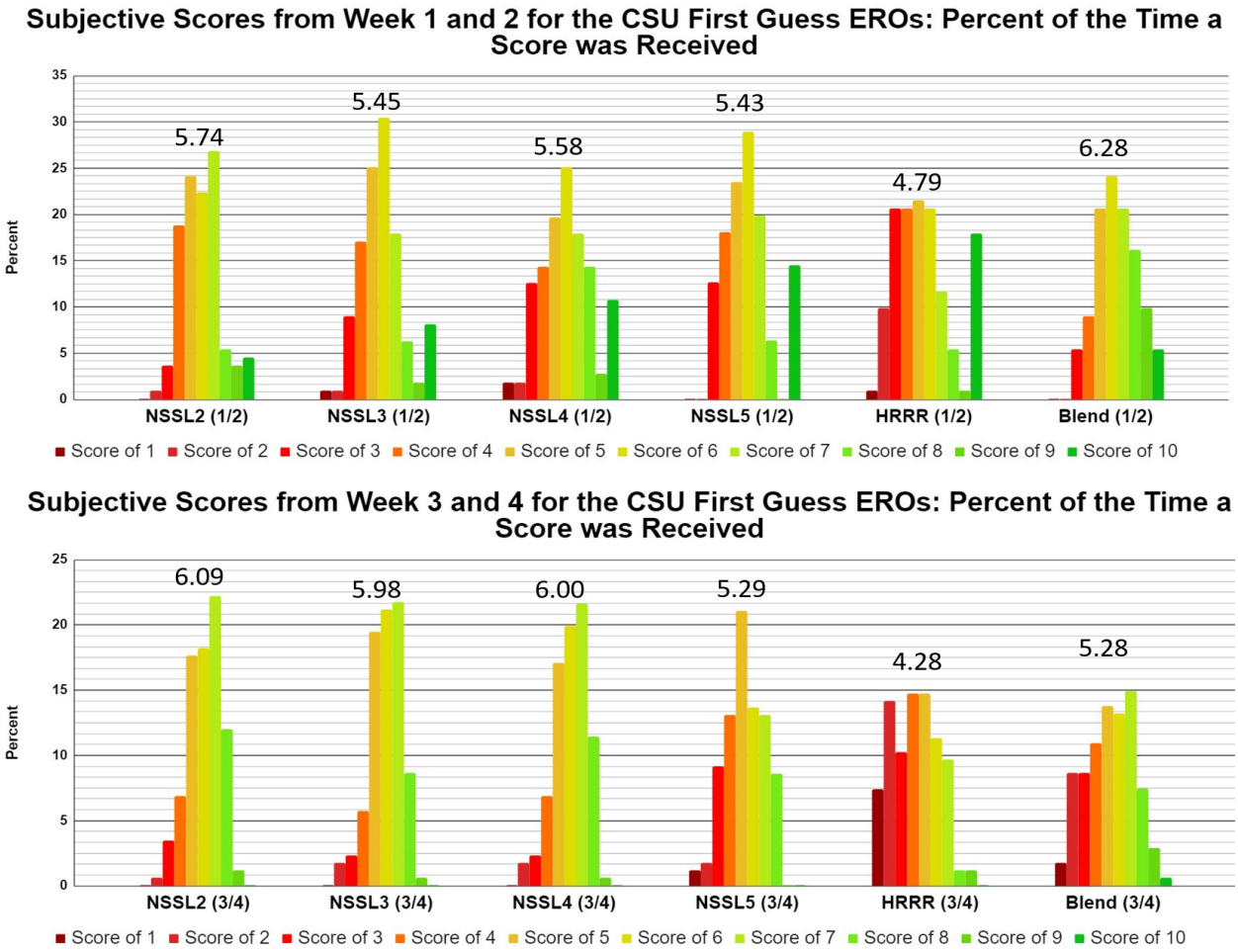


Figure 74: *Same as top image in Fig. 47 but for the subjective analysis of the CSU EROs for weeks 1 and 2 (top) and weeks 3 and 4 (bottom).*

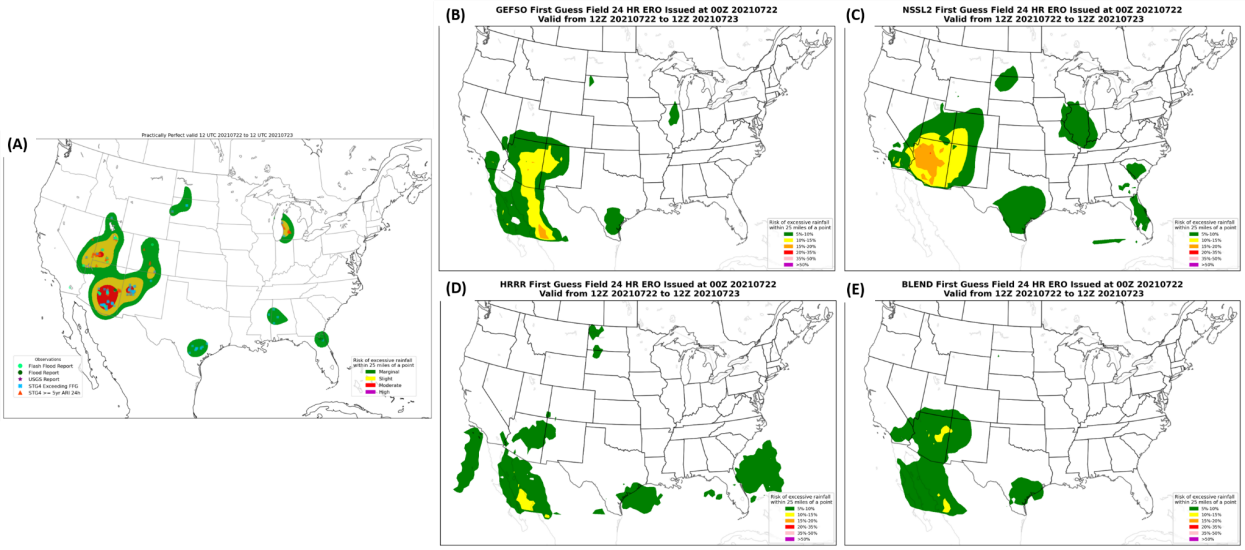


Figure 75: *(A) Practically perfect analysis, (B) GEFS, (C) NSSL2, (D) HRRR, and (E) BLEND EROs valid 12 UTC 22 July to 12 UTC 23 July 2021.*

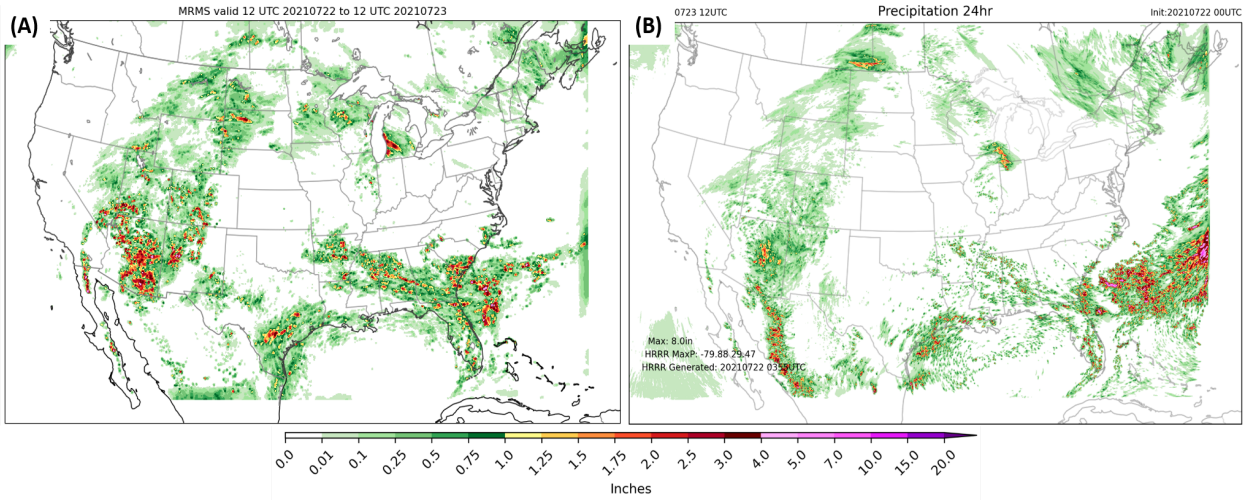


Figure 76: *24 h (A) MRMS QPE and (B) HRRR QPF valid 12 UTC 22 July to 12 UTC 23 July 2021.*

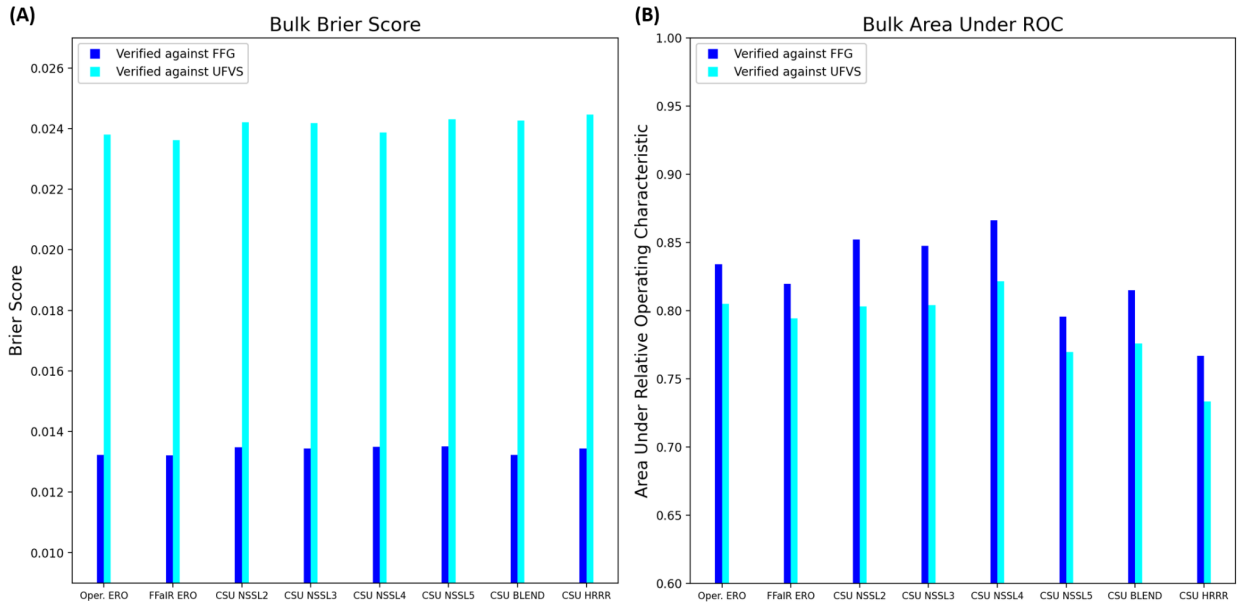


Figure 77: *(A) the bulk Brier Scores and (B) bulk AuROC for the WPC Operational ERO, the FFaIR experimental ERO, and the CSU EROs. Each product was verified against FFG only (dark blue) and using the UFV system (light blue; QPE > FFG, QPE > ARI , and flash flood LSRs, flood LSRs, and USGS gauge reports).*

Despite the NSSL2 being the clear favorite of the CSU "first-guess" ERO configurations, analysis of the BS, AUROC (Fig. 77) and fractional coverage (Fig.80B) suggest the best performing configuration overall was the NSSL4. In fact, when focusing solely on AUROC the NSSL4 also out performed both the Operational and FFaIR EROs, suggesting it does a better job distinguishing events from non-events. However, when comparing it to the Operational ERO via the BSS (Fig. 80A), it generally performs poorer than the Operational ERO, suggesting that even though it can separate out events from nonevents well, it isn't more accurate than the Operational ERO. The same is true for both the NSSL2 and NSSL3 (not shown). Fractional coverage was

one aspect of the forecast that the NSSL4 appears to have performed particularly well, with the factional coverage being well calibrated to WPC's ERO risk categories during FFaIR. It was the most likely of all the CSU EROs to forecast a Moderate Risk during FFaIR (not shown), while the NSSL3 issued no Moderates across the CONUS[44]. The presence of a Moderate Risk in the NSSL4 forecasts and the lack of higher risks from the other configurations were often noted in the participants' comments, especially when it came to NSSL2 and NSSL3. However, it is important to note that there were only a few events over the course of the experiment that were verified via practically perfect as a Moderate Risk.

An example of the NSSL4's ability to highlight the elevated risk both in magnitude and spatial extent over the other NSSL configurations can be seen in Fig. 81. Neither NSSL2 or NSSL3 indicate a Moderate Risk, even though Practically Perfect verification suggests a High Risk was warranted, across northern Missouri. Meanwhile, although the NSSL5 had a small Moderate Risk as well, its extent of the Slight Risk was much smaller than the NSSL4 and missed the magnitude of the flash flooding risk outside of northern Missouri. Although the other CSU "first-guess" EROs performed better than the NSSL4, including the GEFS version that goes into the BLEND, across the region of interest, the NSSL4 was arguably the best for this high impact event.

The results from BLEND were a mixed bag. It, along with the HRRR configuration, had a BS comparable to the NSSL configurations. However its AUROC was the third lowest among the EROs evaluated, likely as a result of the low values from the HRRR ERO. It also appears to be the worst "calibrated" to the ERO risks out of all the options; see the fractional coverage plot in Fig. 80B. On the other hand, the BLEND's BSS (Fig. 82A) was the best amongst the CSU "first-guess" configurations, especially for the first part of FFaIR, prior to the onset of the Monsoon. This suggests that the BLEND was the most "on par" with the Operational ERO when the excessive rainfall risk was more synoptically driven, as it was during the first half of the experiment. This likely is helped by the GEFS ERO which, though not formally evaluated during this time period, historically performs well in such events. Despite the conflicting signal in the metrics, it can be said that the BLEND performed well, had utility, and was liked by the participants overall.

[44] Both NSSL3 and NSSL4 frequently issued Moderates across over water, especially off the southern Atlantic and the Gulf coasts, participants were told to ignore this and the statistics were only calculated over the CONUS.

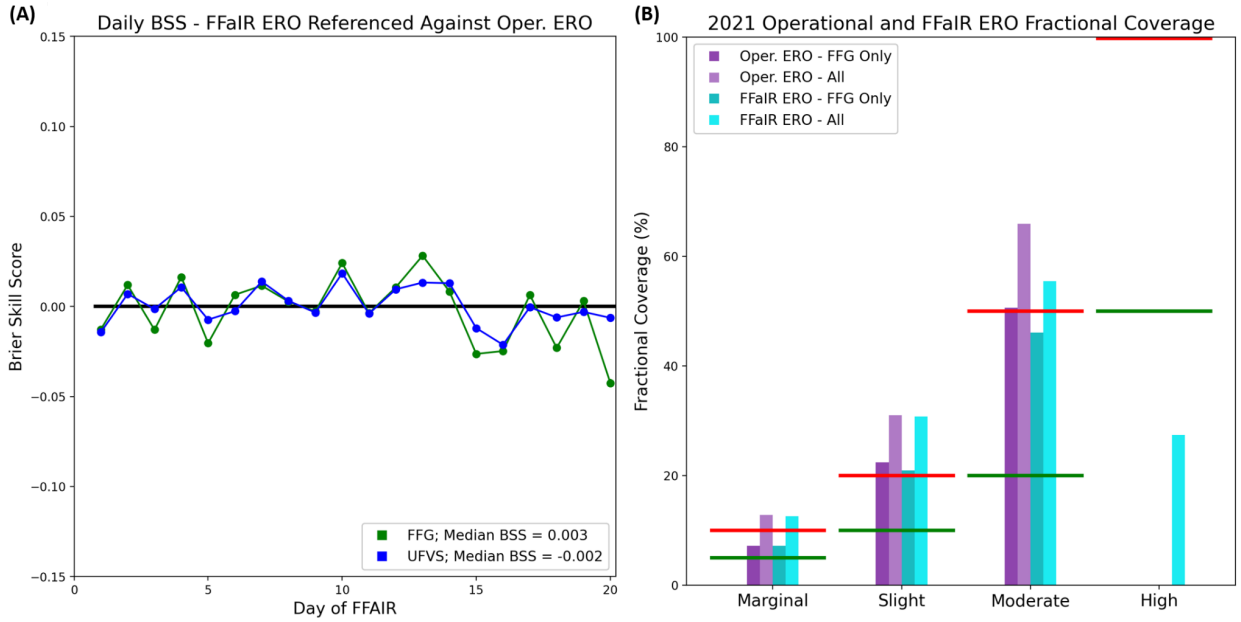


Figure 78: *(A) The BSS for the FFaIR ERO referenced against the Operational ERO. (B) The Fractional Coverage for the FFaIR and Operational EROs.*

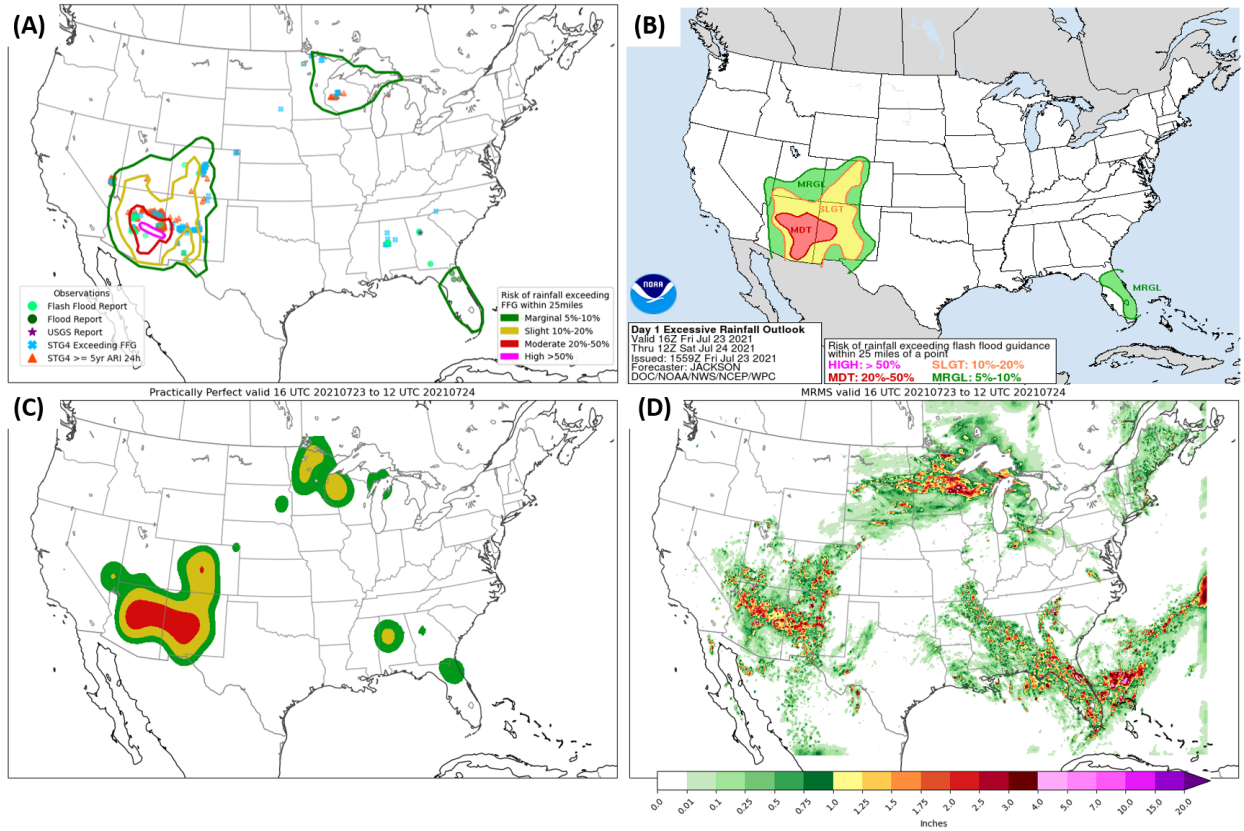


Figure 79: *(A) FFaIR ERO with UFVS observations plotted (see legend), (B) Operational ERO, (C) practically perfect verification and (D) MRMS valid 16 UTC 23 July to 12 UTC 24 July 2021.*

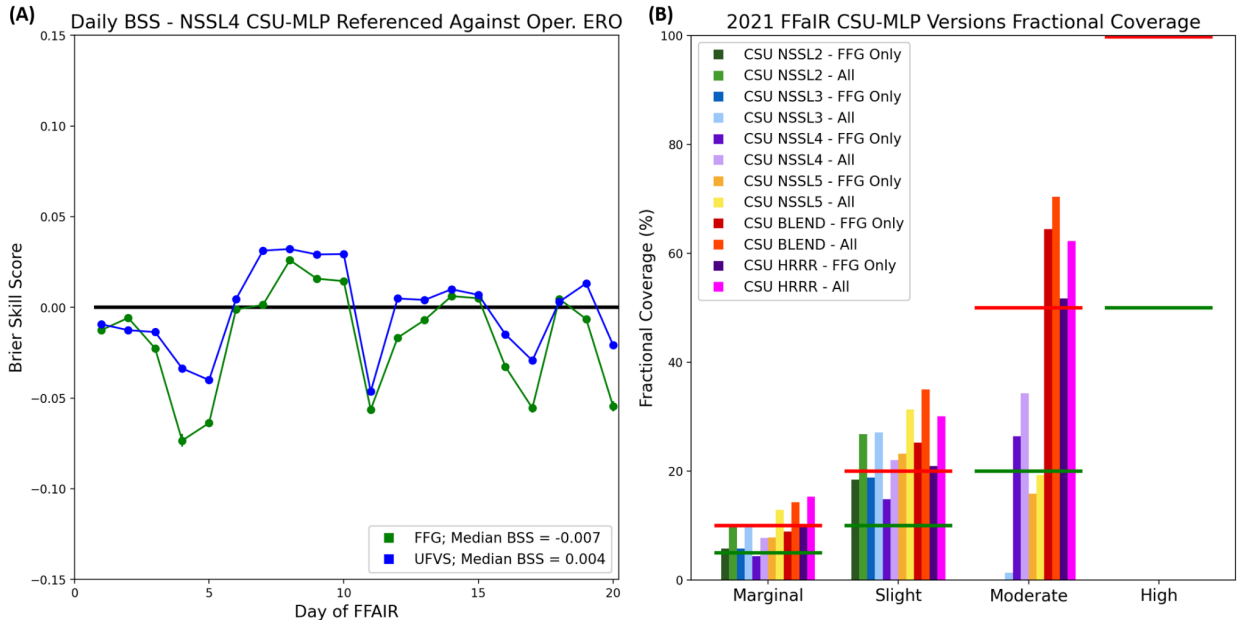


Figure 80: *(A) The BSS for the NSSL4 referenced against the Operational ERO. (B) The Fractional Coverage of the CSU EROs.*

Lastly, the HRRR ERO overall had the worst performance across the verification metrics discussed. This agrees with the results from subjective verification. Again, it appears that the strong Monsoon during the second half of FFaIR helped to drive this poor performance; see the clear decrease in the HRRR BSS in Fig. 82B during the last week of FFaIR, when the Monsoon was the most dominant. The HRRR ERO's inability to identify excessive rainfall risk across the southwest can clearly be seen when looking at the probability of being in a Slight Risk during FFaIR, Figs. 83 and 84. As can be seen in Fig. 83B, the HRRR ERO did not issue one Slight Risk across the southwest. This differs from the FFaIR ERO (Fig. 83A) which had a >30% probability of being in a Slight Risk across portions of AZ and NM and NSSL2-4, Fig. 84A-C, that had probabilities between 24-30%. Even the NSSL5 (Fig. 84D), which was the least likely of the NSSLs to issue a Slight across the Southwest, had a pocket of ~18% chance of being in a Slight risk. The less than stellar performance from this configuration was most likely a result of the short training set used, which was roughly a year of data. The 2020 warm season was part of this training set and as discussed previously the Monsoon was weak last year. Thus the model was not trained to recognize such events. Additionally, a machine learning product can only be as good as the model forecast it is using, thus the HRRR ERO can not be expected to excel when the HRRR QPF in general had a dry bias across the southwest during the experiment. The FFaIR team feels that it would be beneficial to continue development of the HRRR ERO and to retrain the model to include this warm season so an active Monsoon season is in the training set.

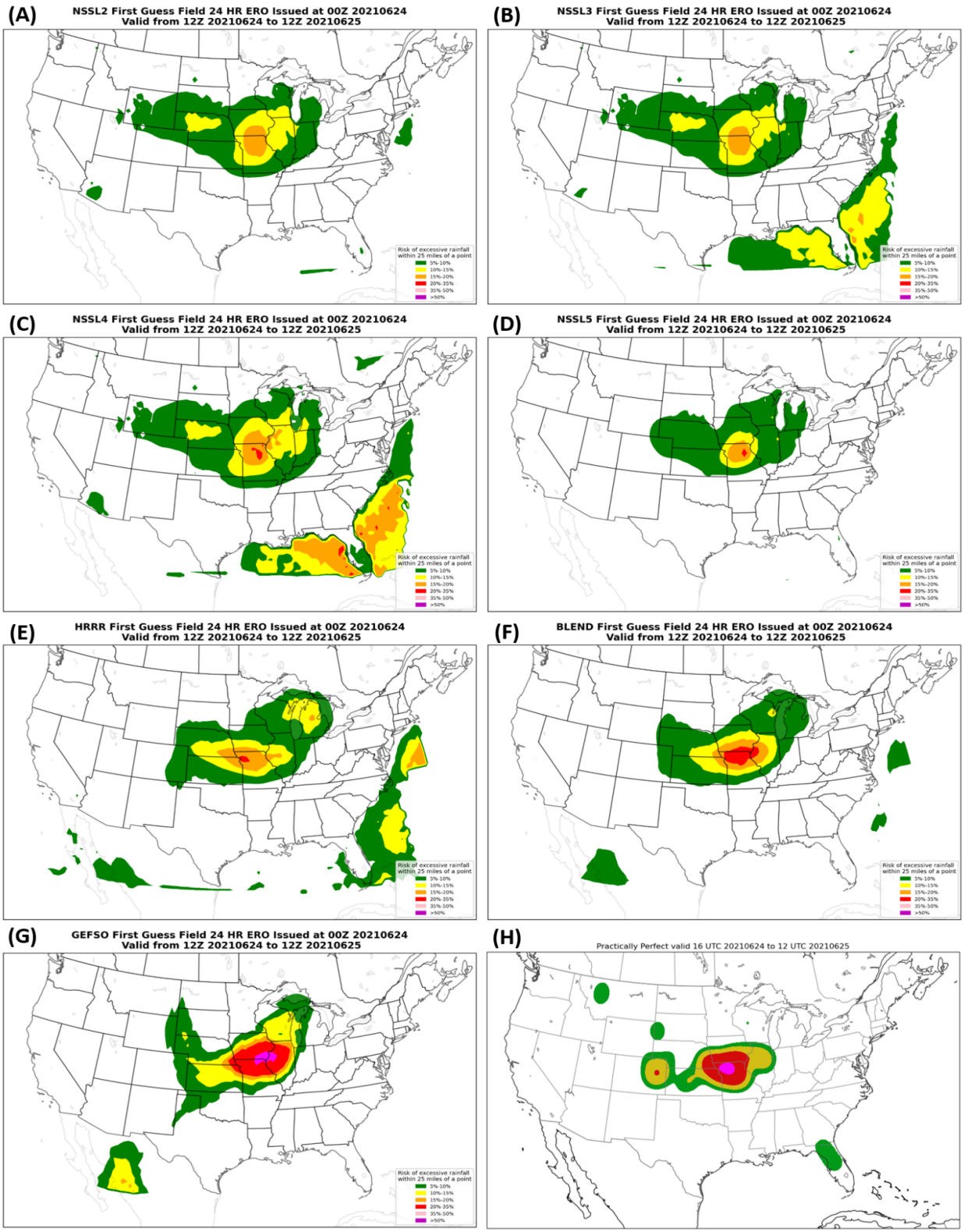


Figure 81: *CSU "first-guess" ERO (A) NSSL2, (B) NSSL3, (C) NSSL4, (D) NSSL5, (E) HRRR, (F) BLEND and (G) GEFS (not evaluated in FFaIR but is operational at WPC and is part of the BLEND ERO) and (H) Practically Perfect verification valid 12 UTC 24 June to 12 UTC 25 June 2021.*

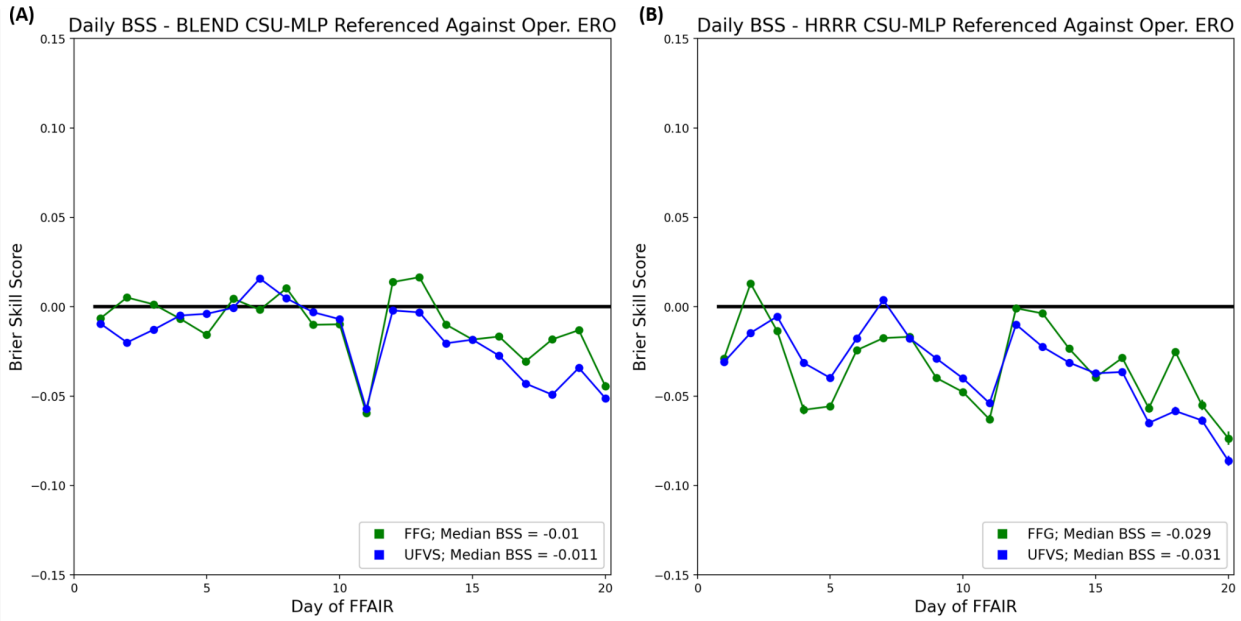


Figure 82: *The BSS for the (A) BLEND and (B) HRRR referenced against the Operational ERO.*

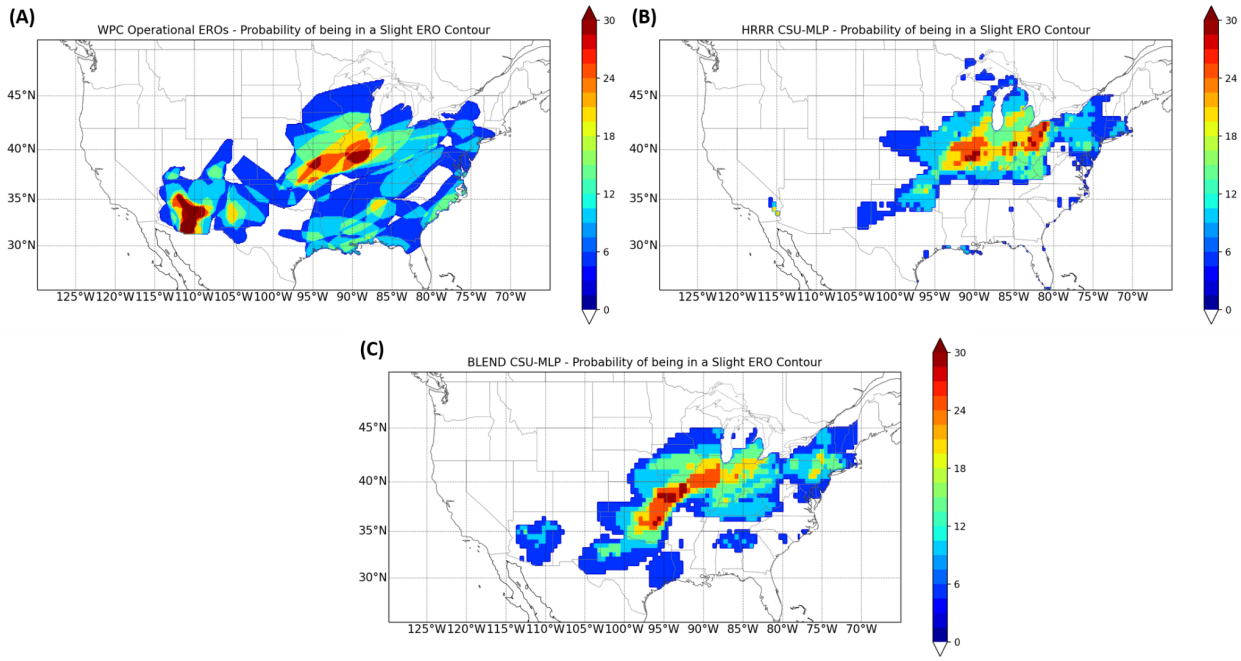


Figure 83: *Probability of being in a Day 1 ERO Slight risk during the 2021 FFaIR Experiment for the (A) FFaIR ERO and the CSU "first-guess" (B) HRRR and (C) BLEND ERO.*

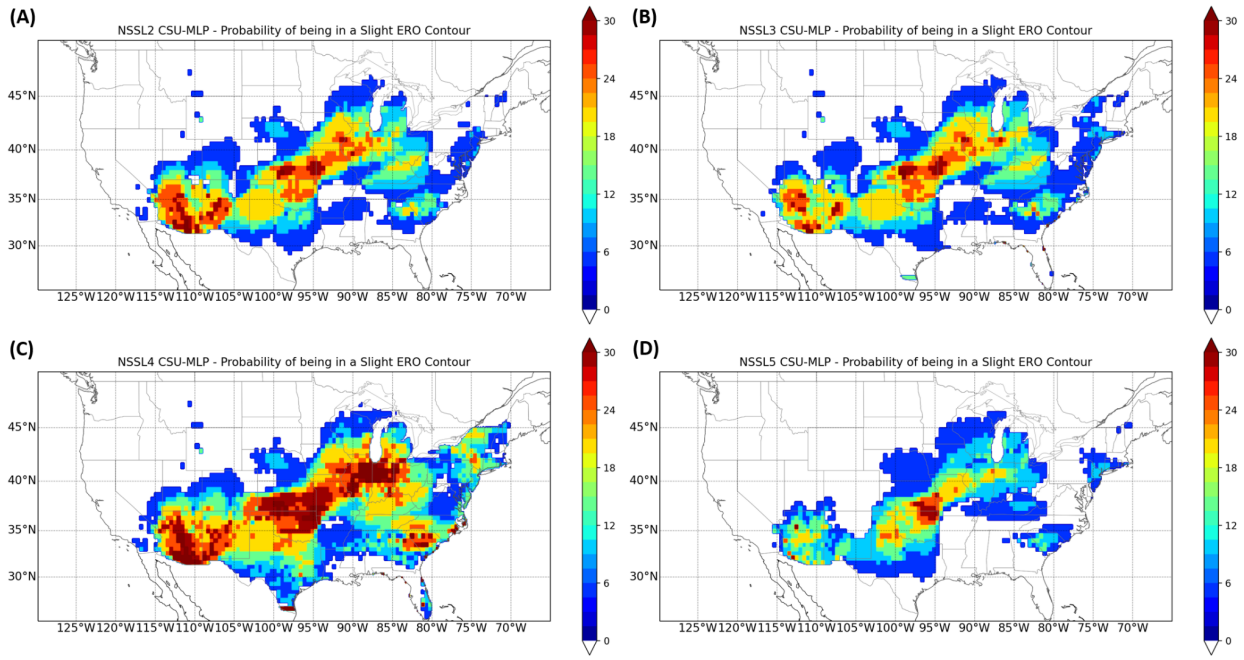


Figure 84: *Probability of being in a Day 1 ERO Slight risk during the 2021 FFaIR Experiment for the CSU "first-guess" (A) NSSL2, (B) NSSL3, (C) NSSL4, and (D) NSSL5 ERO.*

4.3.1.1 Quick Summary

Of the NSSL ERO configurations, the NSSL4 slightly outperformed its counterparts, though in subjective verification the NSSL2 was liked better by the participants. This likely was due to the overforecasing of ERO Risk over the oceans. Participants also noted that they felt like the NSSL4 contours were "less smooth" than the NSSL2. The BLEND ERO was well liked during the first two weeks of FFaIR, having the highest average score of all the CSU EROs during this time period. Again, its shortcomings during the second half of FFaIR across the southwest appear to be a result of the HRRR ERO's poor performance across this region as it struggled to identify the excessive rainfall risk associated with the Monsoon. Overall the FFaIR recommends that the CSU team continues to refine all the CAM-based "first guess" EROs. The products overall provide useful guidance on where excessive rainfall might occur during Day 1. Participants especially liked NSSL2, NSSL4, and the BLEND EROs.

4.3.2 ARI-ERO

Since the ARI-ERO definition focuses on ARI exceedances rather than FFG exceedances, a direct comparison between the FFaIR ERO and ARI-ERO can not be made. Therefore, most of the analysis will focus on the feedback received from the participants about the product. An example of how the ERO and ARI-ERO could differ from one another can be seen in Fig. 85. Note that the traditional ERO is focused across the Midwest and into the Great Lakes region while the ARI-ERO 1 yr ARI extends further to the northwest and north than the Marginal risk of the ERO. Despite the inability for a direct comparison, the subjective scores for the ARI-ERO

are shown alongside the FFaIR and CSU EROs in Fig. 73. As can be seen, the distribution of the ARI-ERO scores are skewed to the right, suggesting that overall the participants felt the ARI-ERO product did well highlighting areas where there was a 75% chance of rainfall exceeding the given 6h ARI thresholds. Before discussing the general feedback about the ARI-ERO product, a brief analysis of the performance of the ARI-ERO will be shown.

A summary of the probability of being in a given ARI-ERO contour can be seen in Fig. 86. Once again, the active Monsoon across the southwest this July can be seen, with there being a >15% chance of being in a 5 yr ARI 6 h contour across portions of AZ, western NM, and southwestern CO. The only other region to see probabilities around this was across MO. Verification was performed by calculating the practically perfect based on ARI[45] exceedances for each recurrence interval (e.g. 1-year, 2-year, etc.), using the same risk definitions (marginal, slight, moderate and high) as those used for the ERO. This resulted in 4 different practically perfect comparisons for verification of each ARI-ERO interval (e.g. compare different practically perfect products that each have 4 risk categories to the ARI-ERO consisting of 4 different recurrence intervals).

Figure 87 shows the ARI-based practically perfect verification for each ARI-ERO threshold. Each figure represents a different practically perfect verification spanning from marginal to high for each ARI threshold. For example, if verifying the 2 yr ARI 6h the practically perfect suggests that during FFaIR, 2yr ARI contours could be justified over a significant portion of the Intermountain West, Central Plains, and mid-West. However, to relate the practically perfect output to the ERO-ARI contours is an open question. Therefore, this information was used to compare each ARI-ERO contour (Fig. 88; green = 1-year, blue = 2-year, orange = 5-year, red = 10-year) to its associated practically perfect based ARI in terms of the ERO risk categories, shown by each group of bars in the plot. In other words, the 1 yr ARI-ERO contour was compared to the practically perfect Marginal calculated using the 1, 2, 5, and 10 year ARIs (the purple fill in each ARI-based practically perfect in Fig. 87), the 2 yr ARI contour was compared to the 4 practically perfect Slights, the 5 yr ARI was compared to the 4 practically perfect Moderates, and the 10 yr ARI contour was compared to the 4 practically perfect Highs. The logic of this approach was to find which ARI-based practically perfect, if any, best corresponded to the ERO-ARI product, with the assumption that 1-year, 2-year, 5-year, and 10-year ERO-ARI corresponded to the marginal, slight, moderate, and high contour of the practically perfect approach, respectively. This assumption has flaws, since recurrence intervals are likely different from the ERO risk categories, but this verification represents a first attempt at evaluating the ERO-ARI.

The bias, FAR, and CSI for each of these comparisons can be seen in Fig. 88. In general, the ARI-ERO underestimates ARI exceedances compared to all practically perfect verifications. The ARI-ERO exhibits the smallest bias compared to the 10-year ARI practically perfect. That

[45] This differs for how practically perfect for the ERO is calculated, which uses the entire UFVS.

said, it is possible that the ARI-based practically perfect over produces instances; Erickson et. al (2021) briefly discusses the shortcomings of using one proxy to create practically perfect verification. Aside from that, it is difficult to identify one ARI-ERO contour that verified "the best" against its respective practically perfect group. Therefore it appears that the ARI-ERO contours are not comparable to the ERO categories and caution should be used when utilizing only one proxy when creating practically perfect. This is not necessarily surprising, since the definition of the contours for the ARI-ERO holds the probability of exceedance constant (75% chance) and changes the value exceeded (1, 2, 5, and 10 yr ARI) while the traditional ERO holds the exceedance constant (exceeding FFG) and changes the probability of exceedance (5%, 10%-20%, 20%-50%, >50%).

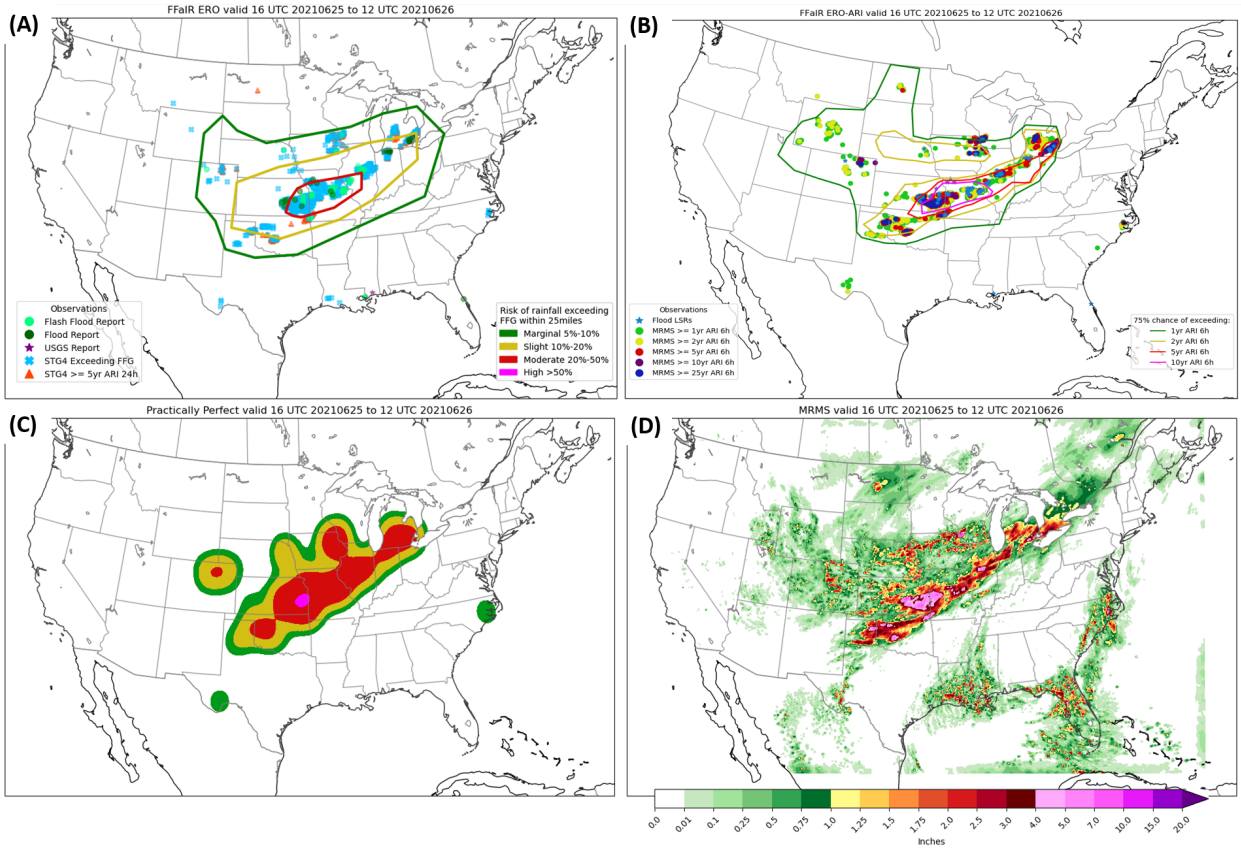


Figure 85: *(A) FFaIR ERO with UFVS observations plotted (see legend), (B) FFaIR ARI-ERO, (C) ERO practically perfect verification and (D) MRMS valid 16 UTC 25 June to 12 UTC 26 June 2021.*

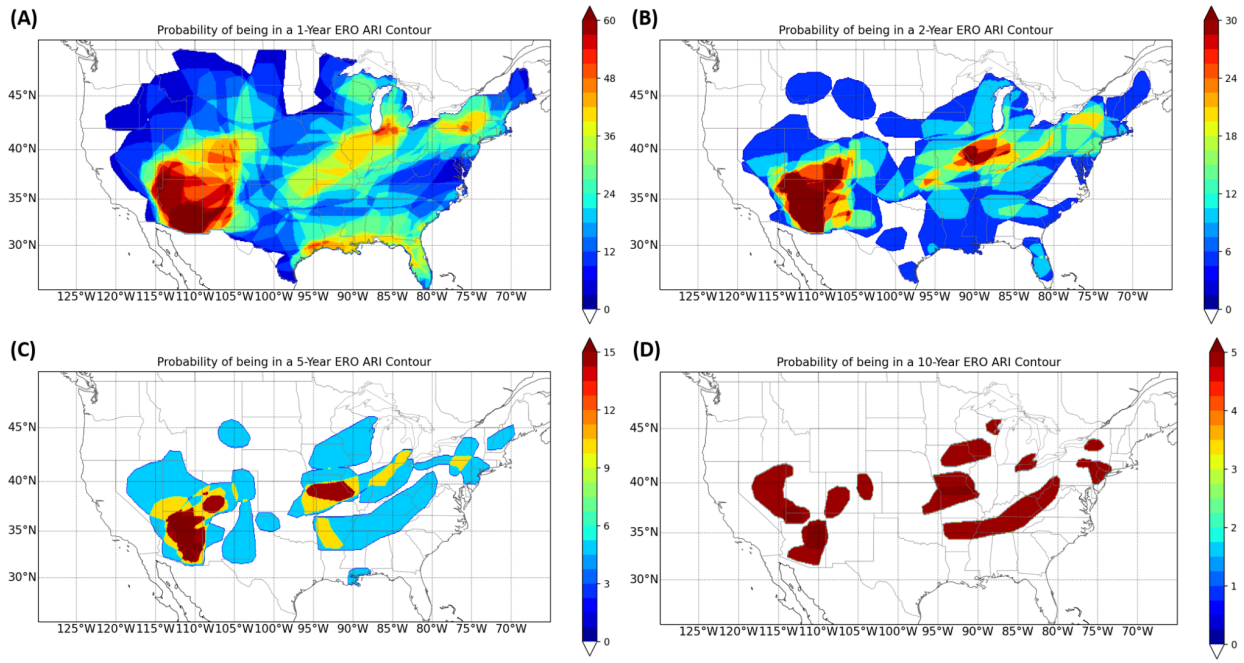


Figure 86: *Probability of being in a Day 1 ARI-ERO (A) 1yr, (D) 2yr, (C) 5yr, and (D) 10yr ARI 6h during the 2021 FFaIR Experiment*

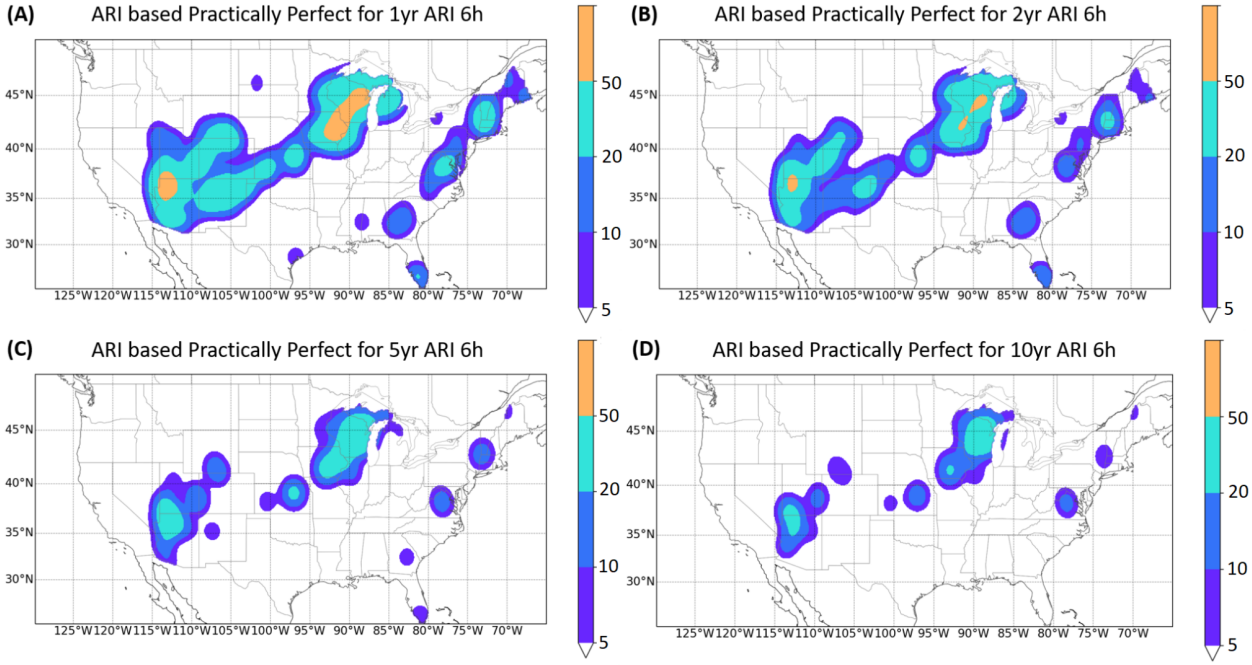


Figure 87: *ARI based practically perfect for (A) 1yr ARI 6h, (B) 2yr ARI 6h ,(C) 5yr ARI 6h, and (D) 10yr ARI 6h. The practically perfect categories are: purple - 5% to 10% , blue - 10% to 20%, green- 20% to 50%, and orange - 50% to 100%.*

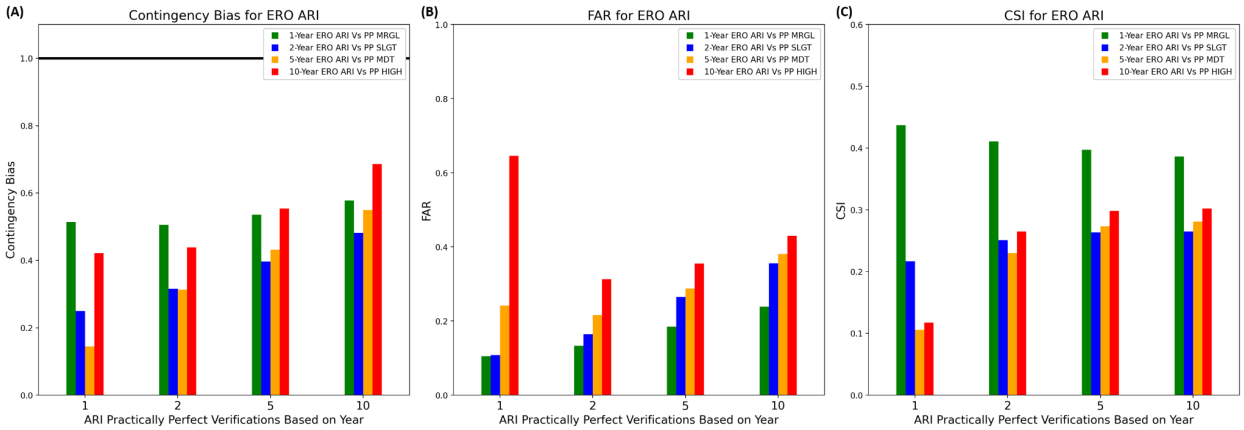


Figure 88: *(A) Bias, (B) FAR, and (C) CSI for the FFaIR ARI-ERO. Along the x-axis are the ARI practically perfect for each year (1, 2, 5, 10). For each year's practically perfect, its respective ARI practically perfect contour (see Fig. 87) is compared to the corresponding ERO risk category: 1yr ARI vs Marginal (green), 2yr ARI vs Slight (blue), 5yr ARI vs Moderate (orange) and 10yr AIR vs High (red).*

Focusing now on the feedback from participants about the ARI-ERO product, as a whole, the general feedback was positive. For the most part participants liked the idea of the product but noted it had some shortcomings. One common comment was that they felt that starting with an 1yr ARI exceedance was too low, with some even commenting that 2yrs might be too low as well. Many suggested starting at the 5yr ARI. They also noted that the product seemed more like a heavy rainfall probability product than a flooding product, often stating that exceeding a higher ARI does not always translate to flooding. These two sentiments are summed up in the following comment: "I really like the potential of the ARI ERO to inform on the threat of heavy rainfall. I think the higher ARI values of 5, 10, and 25 years are more useful on a regional or CONUS scale. The ARI ERO could perhaps be one of many ingredients in assessing flash flood risk, but obviously there are other important factors such as antecedent conditions and land use classification."

The latter portion of the previous comment, stating that using an ARI product as guidance to help create a flood risk product (in this case the ERO), or to use it in tangent with the ERO to identify risk was another sentiment that was common throughout the feedback from participants. Some additional examples of such comments are:

- "I think it is definitely beneficial and helpful to the ERO to take the ARI into account. It really did seem to make a difference, even with the uncertainty of location from model to model."
- "I like ARIs for identifying initial threats, but then more info is needed to finalize FF threats based on those forecasts"
- "I think one element that's particularly helpful with ARI is being able to contextualize a flood magnitude potential. Speaking in terms of ARI helps

with messaging, especially when focused on a singular flood episode in the warning environment."

- "It is very useful to highlight anomalous events. It is helpful to identify heavy rain that may not result in flooding but still have impacts (water ponding on roads, river rises). Without the dependence on FFG, the local forecast office can analyze antecedent conditions to see the impact of the heavy rain on flooding. A lot of our impacts are a result of high rain rate for 2-3+ hours regardless of flash flood guidance so the ERO ARI ties nicely into that. The traditional ERO has it's place too. I like the idea of having both to assess how each can be useful. We use 1 hr, 3 hr, and 6 hr ARI in FLASH during heavy rain events so the ERO- ARI and the guidance driving it could be a great situational tool 1-3 days before an event."

An additional shortcoming of the product, aside from starting with 1yr ARIs, that was discussed during both the creation of the product and the verification was the difficulty of not applying the ERO risk categories to the ERO-ARI. Meaning, participants often wanted to call the 1yr ARI a Marginal risk, the 2yr ARI a Slight risk, and so on. Participants noted that the difficulty of separating the ERO risks from the ARI-ERO product resulted in hesitation in drawing contours for the 5 and 10 yr ARIs. There were times during the collaborative drawing process for the ARI-ERO that participants would say things like "I am not sure that warrants a High risk" and they would need to be reminded that drawing for the 10yr ARI does not equal an ERO High risk. One factor that likely led to this struggle was the fact that the same color scale (green, yellow, red, pink) was used for the ARI-ERO thresholds as are used for the ERO risk categories. Another hindrance was likely in the name of the product itself. Having ERO in the name likely gave the participants a preconceived idea of what terminology would be applied to the product. This was confirmed in the end of the week survey, where numerous participants suggested changing the name as to not confuse the two products.

Based on the feedback from the participants, the FFaIR team will likely continue with the ARI-ERO exercise in next year's FFaIR experiment. However the product itself will likely be renamed, as suggested by participants. It will also be refined to address some of the shortcomings identified by the participants.

4.4 MRTP

The Maximum Rainfall and Timing Product (MRTP) was designed to have all participants draw multiple rainfall contours in a 6 hour period in addition to drawing an area of six hourly maximum precipitation rate above 1"/hr. Of the 20 FFaIR experiment days, 18 had at least 1 Flash Flood Warning for the domain of interest. Five of the 20 days were located in the Southwest US for the Monsoon with the remainder east of the Rockies for various mesoscale flooding events and mesoscale convective systems (Fig. 89). According to our participants using the event driven surveys (Fig 90), for 13 days the median flooding potential was above 70%

while for 12 days the median flood damage potential was above 50%. Only 4 of the 20 days had an observed 1” areal coverage above 100k km^2 while 4 days had rainfall maxima above 6” (Fig 90E). Thus the events observed during FFaIR tended to be smaller, less intense (with few MCS events), but still relatively impactful.

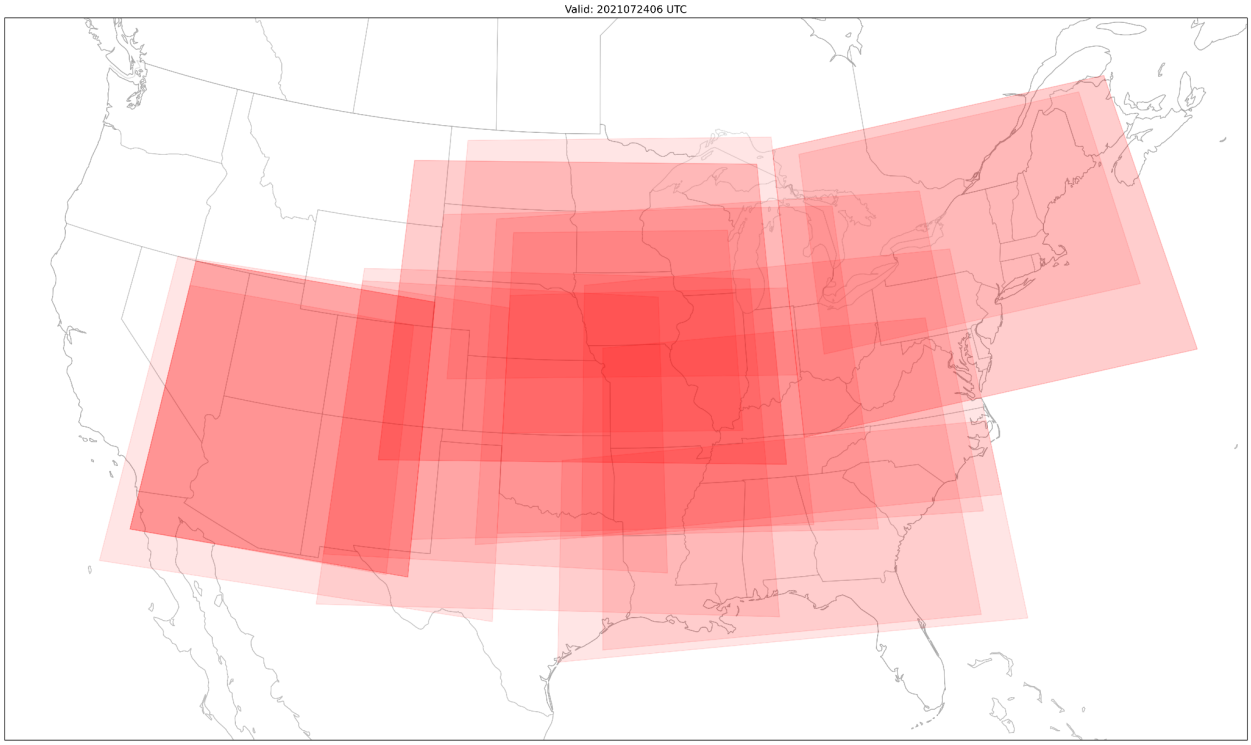


Figure 89*: Bounding box of each MRTP forecast domain for each of the 20 experiment days.*

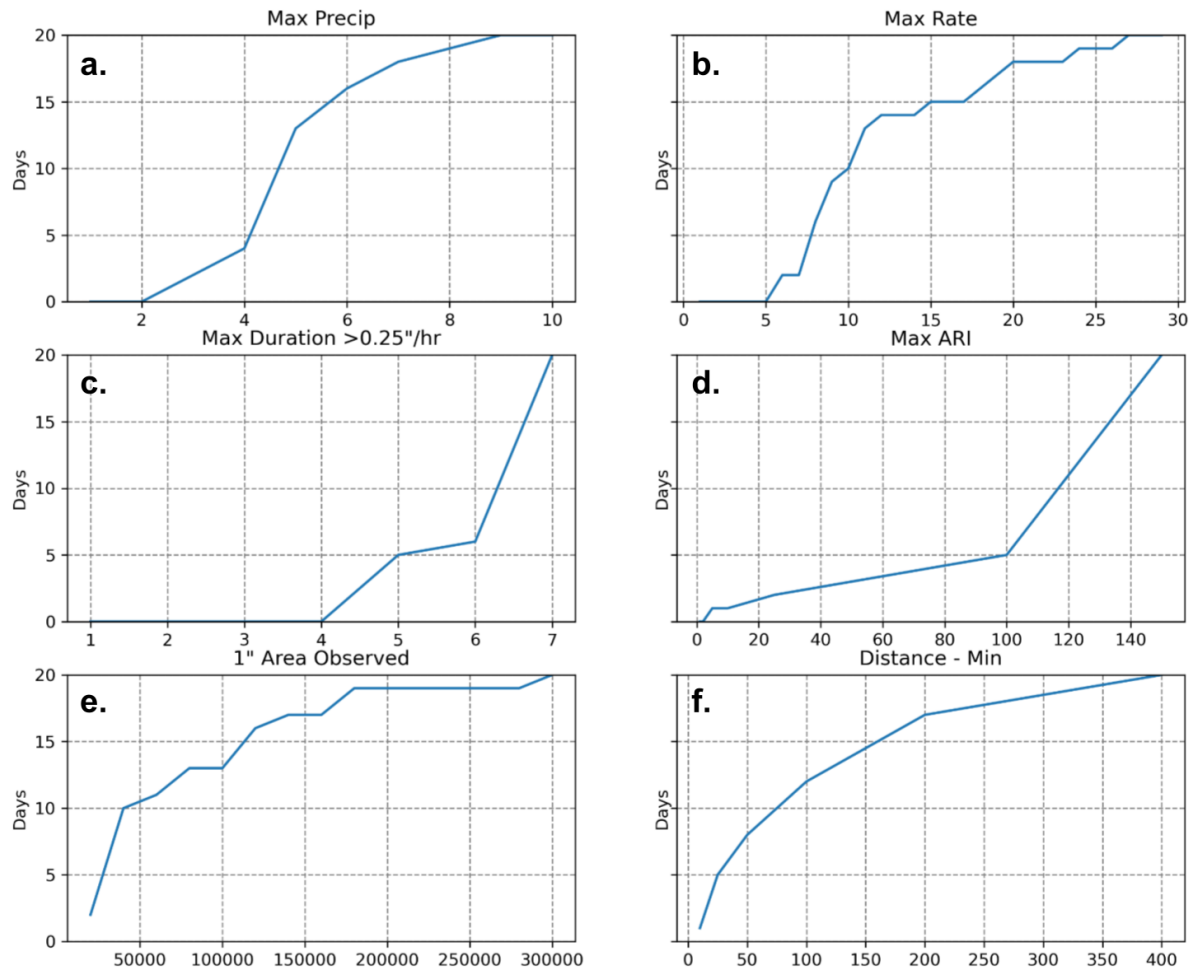


Figure 90*: The occurrence frequency by forecast days for MRTP domain (a) Maximum Precipitation, (b) Precipitation Rate, (c) Duration above 0.25"/hr, (d) ARI, (e) 1" Area, and (f) forecasters minimum distance for the location of the maximum rainfall.*

4.4.1 Performance Diagrams

Verification of 1" areal coverage for each of the 20 days was performed for all participants, and as many CAMs (typically 13) as possible per the valid time of the forecast (as many as 12 cycles). Examining all 20 days (Fig. 91), participants tended to be well distributed in POD and highly clustered in SR on the majority of days; i.e. those days with low 1" coverage (<30k km^2) and thus relatively low CSI. On 5 days, clustering with regard to POD increased; i.e. days with relatively high coverage and as such CSI increased substantially. On these higher coverage days the best forecasters were competitive in CSI with the best models. Typically the best models were models/cycles the forecasters never had access to during the experiment because they were available within 6-12 hours of the event and after the MRTP was due. Participants had primarily operational models after 12 UTC and experimental models, which were usually just the 00z cycles, prior to the forecast activity that ended around 21 UTC each day.

On most days, with most models/cycles, model bias fluctuated from around 1 on the highest coverage days to greater than 1 on most low coverage days. While our MRTP domains were by no means small, there could be many areas of precipitation to forecast for, with many different mechanisms operating. Therefore, while the bias appears around 1 frequently this is not to suggest that model forecasts were of high quality.

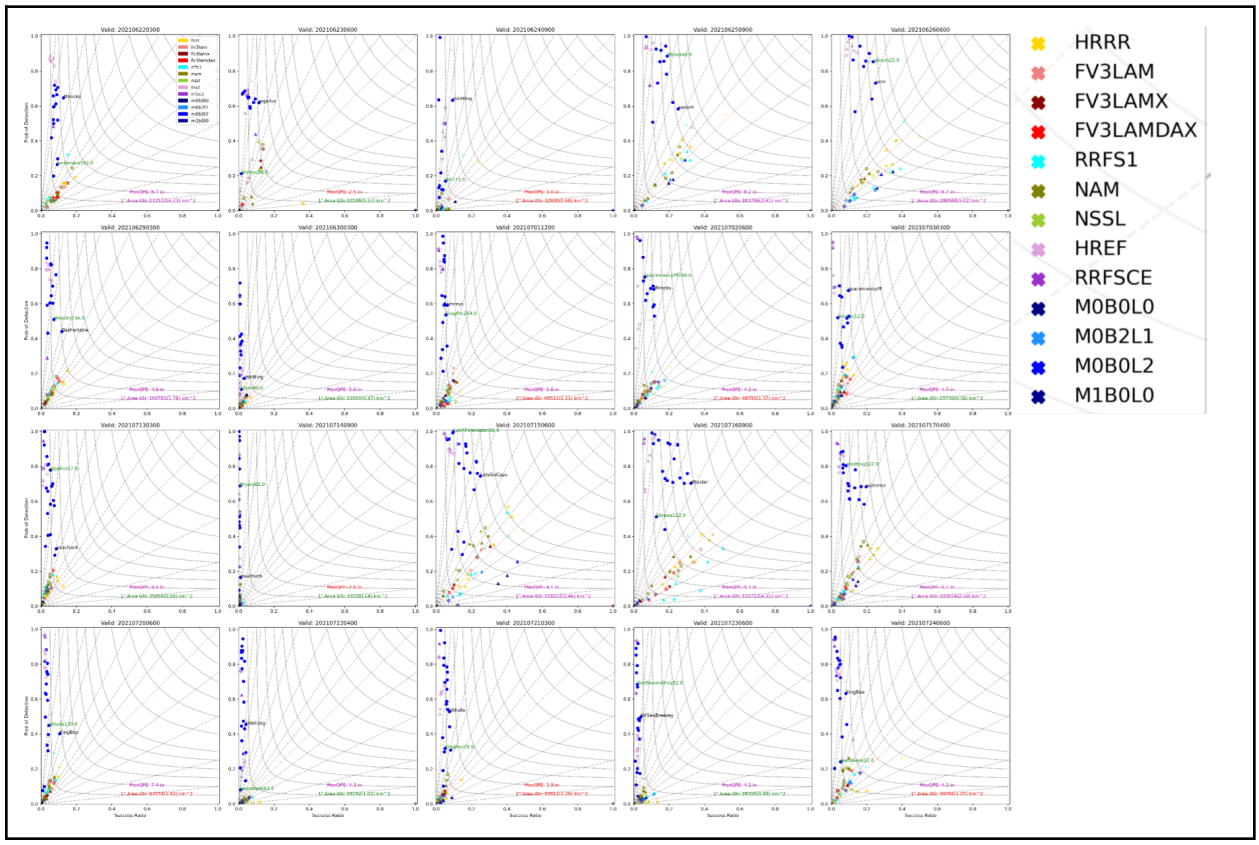


Figure 91: *Performance diagram for all 20 days of FFaIR, arranged by week (top to bottom) and day of week (left to right). Participant forecasts are in blue dots with each day's best 1" CSI and minimum distance to observed max highlighted. Models have their own color key, with symbols representing each model cycle.*

Broken down by week or aggregate over the whole experiment (left and right images in Fig. 92 respectively), participants tended towards a bias of around 8, reflecting the technique of drawing their 1" contours large so as to capture as many events as possible. Given some of the large uncertainties that our collection of experimental models provided, participants chose a strategy that minimized effort but focused attention on the higher thresholds. Subsequent precipitation thresholds were much smaller and reflected a disconnect in our experiment design, namely that the 1" contour was used for convenience not skill, which was something participants discussed in every verification discussion. While participants stated they were worried about the false alarm area, that still wasn't enough motivation to reduce their extent of the 1" contour. This tradeoff was discussed at length for what they were trying to accomplish via performance metrics and the main takeaway was that POD was just more important than FAR.

The models' typically have a bias just greater than 1. Upon daily visual inspection, model forecasts seemed too sparse and incoherent, so this result of a bias ~ 1 is different than expected. The models were finely detailed compared to participants and so it may be that models appear good with respect to metrics, yet their overlap with observations may be sparse and/or incoherent. These visual depictions improve as we get to higher areal coverage events and models may miss specific smaller areas and nearly capture the large areas. This different tradeoff in the models may give the impression that participants were much better than the models but in terms of quality (e.g. CSI) they are evenly matched.

By far the best models were operational (HRRR followed by the NAMnest). While we focused exclusively on events after 21z and particularly after 3z, few of the experimental models had a diurnal cycle of 6h precipitation that came close to the MRMS verification dataset used here. Most experimental models had a rapid dropoff in 6h precipitation after 00 UTC no matter the synoptic pattern with perhaps the exception of a tropical cyclone. Thus the experimental models were disadvantaged in this particular activity despite having had very few MCSs during the 20 days.

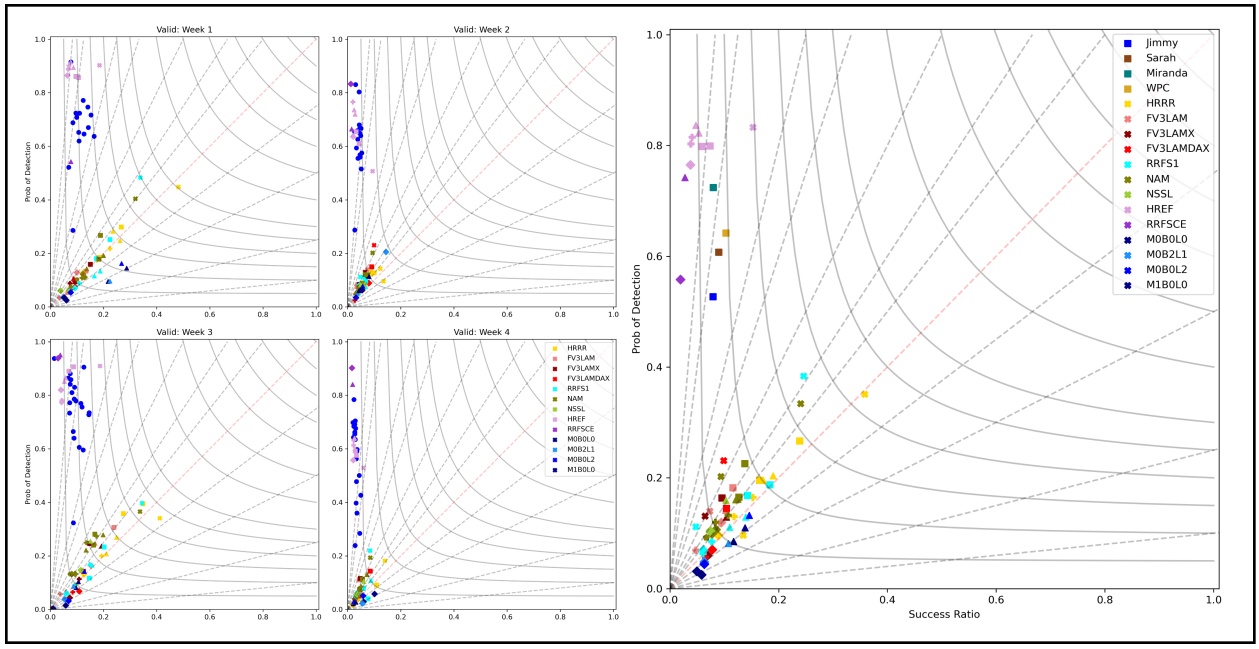


Figure 92: *Same as Fig. 91, except for each of the 4 weeks (left) and the aggregate over the whole experiment (right) for select participants.*

4.4.2 MRTP Case Study of 25 June 2021

The 25 June 2021 case was discussed briefly in Section 3, and had MRTP valid from 03 to 09 UTC. This was the 3rd highest areal coverage (186k km^2) event and highest maximum point accumulation (8.2") during FFaIR. On the synoptic scale, a 500 hPa ridge axis was in place over Nebraska and Iowa as an overnight MCS propagated from northwest to southeast into MO, leaving behind a weakening outflow boundary as a cold front from MN sagged southward. These

boundaries eventually intersected over northeast KS serving as a focus for convection initiation of the overnight MCS that we were forecasting. As this MCS propagated southeast, the low level flow was perpendicular to the outflow boundary in MO and could thus serve as a warm front where stationary or training convective elements on the southwest flank of the MCS would increase precipitation totals. Many of the models indicated the potential for maximum accumulations of 6-10" of rain, in a band along the outflow. The inconsistency all models depicted was whether the maximum precipitation area would be along the outflow or further to the east along the northern edge of the MCS (in the likely location of a bookend vortex or MCV). Individual models flip flopped between cycles indicating either of the above mentioned scenarios, which they consistently did through 3 cycles.

The 6h, full CONUS verification for the HRRR, where forecasts were verified at +/- 2 hours of the valid time, indicated that the HRRR performance varied according to cycle time. For instance, for the HRRR's 00 UTC 24 June 2021 cycle, the 1" verified better at a later forecast hour (i.e. the 11 UTC 6h precipitation had significantly higher skill than that at 09 UTC) while for the 06 UTC cycle, scores were relatively even across all 5 times verified (Table 5). The 12 and 18 UTC initializations depicted significantly better forecasts earlier than the valid forecast hour for most thresholds. Despite this apparent timing issue, analysis of the HRRR showed no strong bulk preference for timing errors (specifically at 00 UTC initializations; not shown). However, later forecasts generally had improved performance only minimally above the target verification hour up through 36 hours of forecasts. The RRFS1 (Table 6) had similar performance (in terms of timing) as the HRRR and generally comparable CSI to the HRRR for most evaluation periods and thresholds. However, like any verification statistic, the spatial patterns of the forecasts were different relative to the initialization time. Some spatial patterns had better correspondence with observations, with the domain maximum in very different locations with at least one initialization coming close to that observed.

Overall, model performance was particularly good, as this was a synoptically evident event. One way to understand how well the model forecasts were is by comparing their results from this event to the daily distribution of CSI scores obtained from May 21-July 31st 2021 over the CONUS for each 00 UTC cycle and forecast hour. This creates an abbreviated model climatology to get a sense of how well the model performed. Focusing on the HRRR performance (Fig. 93), the regional MRTP statistics for this event (left) are compared with the abbreviated CONUS climatology results (right). Such a comparison is applicable assuming that the bulk of the 1" rainfall fell over the MRTP domain, and that criteria is met for this case. The regional and national CSI are consistent and are well above the 90th percentile of days for both the HRRR and RRFS1 00z cycles (indicated by the arrows and color coded asterisks) for the particular 6h periods at their respective valid forecast time.

Table 5: *Tabular listing of CSI values from the HRRR by precipitation threshold for each forecast valid time across the 4 cycles (00-18 UTC) on 24 June 2021. Red highlights indicate the highest CSI for that threshold across the forecast valid times, while the red bold values indicate the highest CSI for each threshold.*

HRRR/ INIT (CSI)	00				06				12				18			
Thresholds (inches)	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**
07	.17	.14	.06	.01	.24	.15	.05	.01	**.36**	.25	.13	**.07**	.34	.26	.12	.04
08	.21	.18	.09	.02	.26	.16	.04	.01	.34	.22	.11	.05	**.36**	**.27**	.08	.01
09	.26	.21	.11	.03	.27	.15	.02	0	.30	.17	.07	.01	**.36**	.24	.03	.0
10	.30	.24	.12	.04	.28	.16	.02	0	.25	.12	.03	0	.33	.2	.02	.0
11	.33	.26	**.15**	.03	.28	.18	.03	0	.19	.08	.01	0	.29	.15	.02	.0

Table 6: *Same as Table 5, except from the RRFS1 model.*

RRFS1/ INIT (CSIx100)	00				06				12				18			
Thresholds	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**	**.5**	**1**	**2**	**3**
07	.18	.14	.08	.05	.25	.23	.10	.04	**.30**	.19	.02	.00	**.30**	**.26**	.15	.05
08	.21	.17	.09	.05	.27	.24	.13	.08	.28	.17	.02	.00	.29	.24	.12	.03
09	.23	.19	.09	.04	.27	.21	.15	**.10**	.24	.13	.02	.00	.27	.20	.08	.01
10	.26	.21	.11	.04	.24	.18	**.16**	.09	.20	.12	.01	.00	.24	.16	.03	.00
11	.26	.22	.12	.04	.21	.16	.11	.03	.16	.07	.00	.00	.22	.13	.01	.00

Regional MRTP Domain Statistics

Cycle	Model	CSI
20210623 12z	HRRR (RRFS1)	0.16 (M)
20210623 18z	HRRR (RRFS1)	0.17 (0.03)
20210624 00z	HRRR (RRFS1)	**0.21** (0.18)
20210624 06z	HRRR (RRFS1)	0.12 (0.19)
20210624 12z	HRRR (RRFS1)	0.14 (0.11)
20210624 18z	HRRR (RRFS1)	0.2 (0.17)
20210625 00z	HRRR (RRFS1)	**0.36** (0.23)

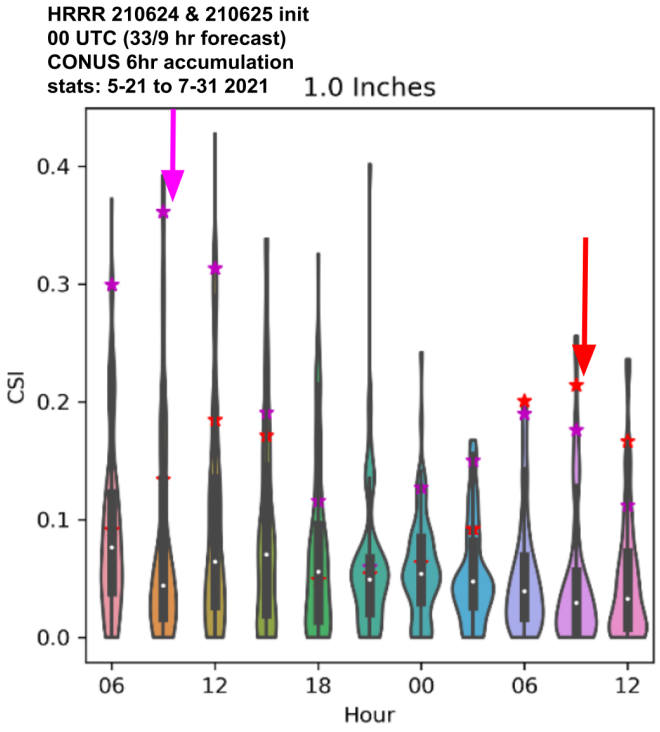


Figure 93: *A comparison between the HRRR and RRFS1 CSI values by initialization time valid for 09 UTC 25 June 2021 (left). Violin plot (right) of daily HRRR CSI values from the 00 UTC initializations every 3 forecast hours, with the red (purple) stars highlighting the HRRR performance from the 00 UTC initializations on 24 June 2021 and 25 June 2021, respectively. HRRR CSI values in red (f33) and purple (f9) are referenced via similarly color coded arrows for their respective forecast hours.*

The NAMnest and HRRR, which had fluctuating performance on the day of forecasting, were still highlighting similar areas, though bouncing around north and south of where precipitation was eventually observed. Performance generally increased from the 23rd into the 24th with a final jump for the last available cycle on the 25th. Participants were relatively consistent in the location of precipitation and thus improved over the fluctuating guidance the day of, and more specifically from Day 2 to Day 1. The largest improvements for Day 1 were a reduction in number of participants with large area 1" contours in western NE and an increase in 1" areal coverage across the front the KS. This change as a participant ensemble can be seen in Fig. 94 and examples of how the change was seen individually can be seen in Section 3, Fig. 17.

As previously shown Fig. 91 (fourth PD from the left in the top row), the participant with the highest CSI correctly forecast the area with a minimal bias (~2.5) relative to other participants and slightly above that of the best models available to participants (~1.25). The overall shortest distance to maximum rainfall was 6km by "MIRocks". Eight of the 12 participants were in the same band of CSI (.15-.21) as all of the available guidance, and visual inspection of all spatial forecasts indicated that participant forecasts focused on the flooding event as depicted by the ensemble of participant forecasts across all thresholds drawn, (Fig. 95).

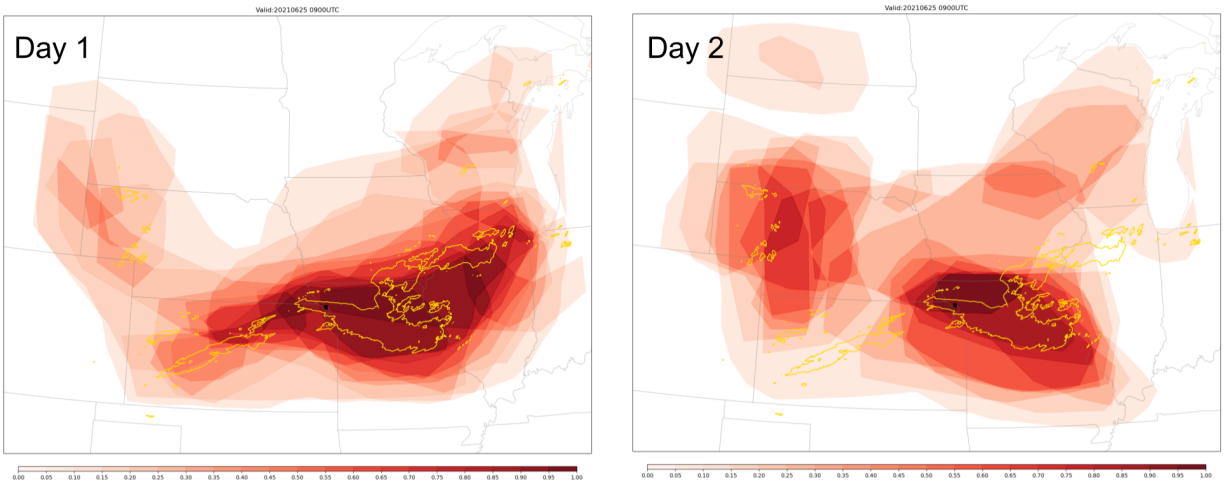


Figure 94: *Participant Day 1 (left) and Day 2 (right) ensemble forecast for 1 inch for 09 UTC 25, shaded in increments of 10%.*

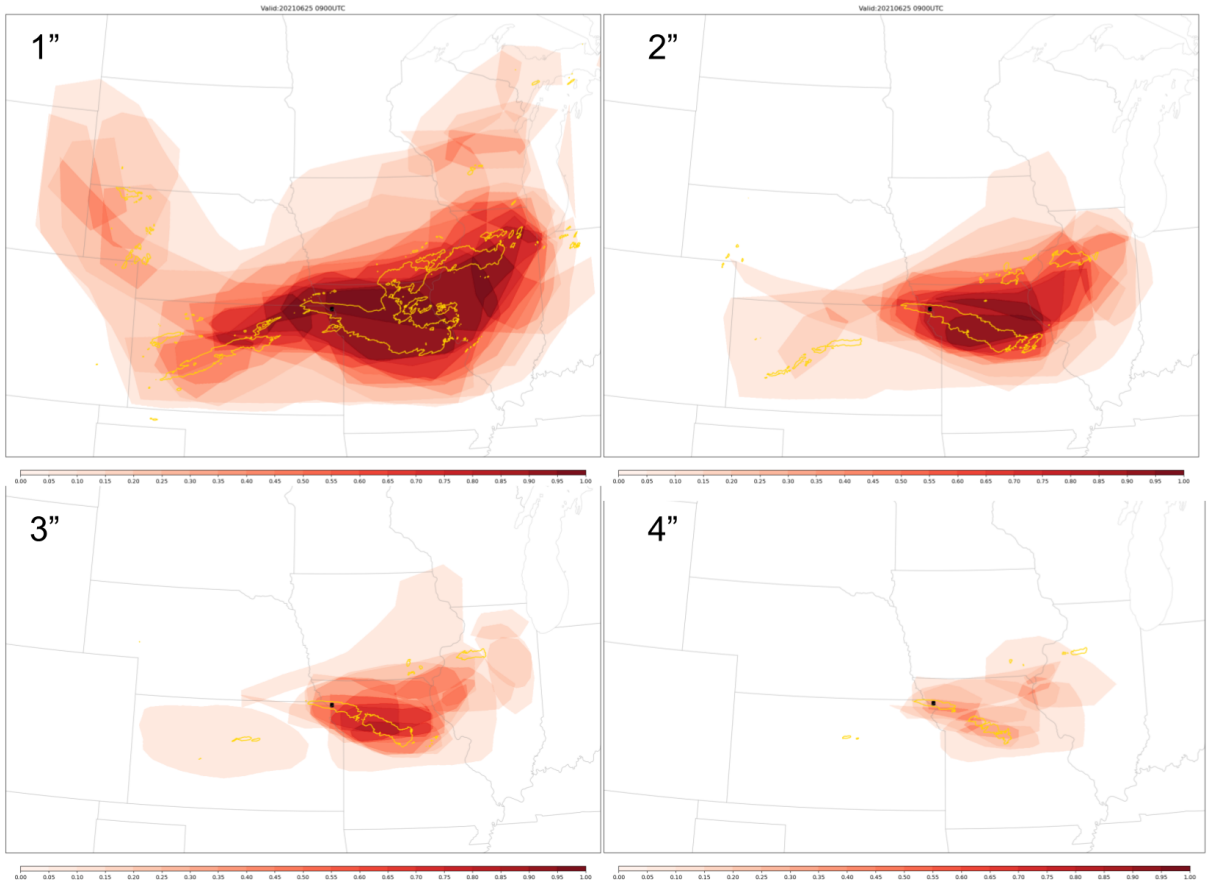


Figure 95: *Similar to Fig. 94, but for the Day 1 precipitation thresholds: 1 inch. 2 inches, 3 inches, and 4 inches.*

4.4.3 Participant Evaluation of MRTP: Survey Results

4.4.3.1 Used, Useful and Usable

Each participant for the MRTP experiment was assigned a model or ensemble that they were instructed to view. It was their choice to accept or pick a new model to base their forecast on and record in our survey (full survey can be found in Appendix C) which models they found useful and which they didn't. Models could be assigned as many categories as they wanted. As can be seen in Fig. 96, by far the most "used" models were operational (HRRR, HREF, NAM) as they are the most familiar. But almost no experimental model garnered dominant attention above the option "considered". The previous problem mentioned with the diurnal cycle of precipitation was noticed in multiple ways and is reflected thusly. Participants usually mentioned the lack of precipitation areas in many different locations while the primary area of precipitation would be smaller than they expected and when compared to the operational guidance.

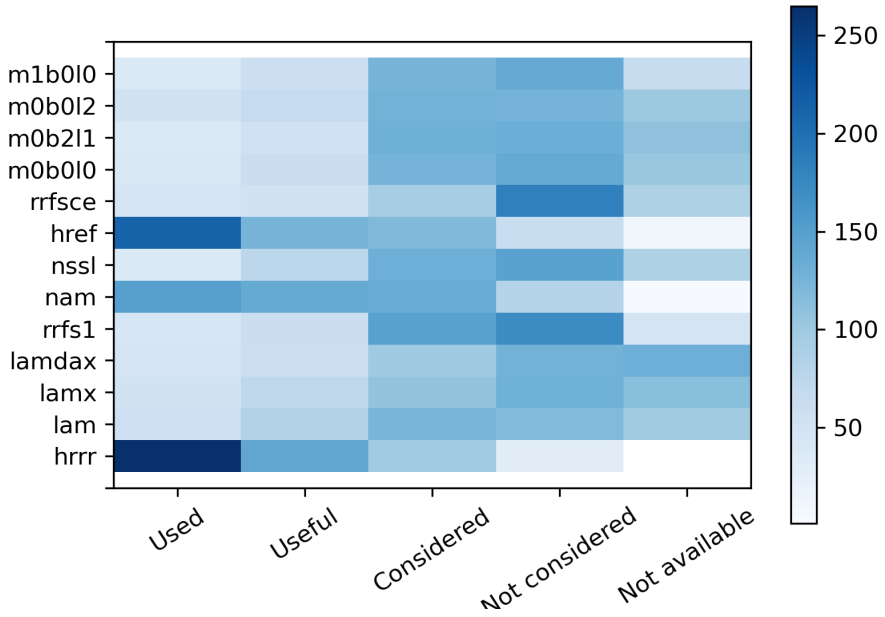


Figure 96: *A heatmap of the counts of the models available to participants for use in the MRTP activity according to how participants utilized the model on any given day (Used, Useful, Considered, Not considered, or Not Available).*

4.4.3.2 Flooding

One objective for this year was to begin codifying the risk of flash flooding and the factors that may contribute to it. We added survey questions that asked about the probability of flooding and the flood damage potential. Synthesizing participants' expectations in this way, we wanted to see if there was any relationship between maximum rainfall and the other quantitative measures we collected with their perception of the probability of floods or damages.

Looking at Fig. 97 it can be seen that the participants were skewed lower when predicting flood damage versus flooding by at least 10-15% on average. As mentioned earlier, there were

very few events where the majority of participants felt flood damage was highly likely. A selection of 6 cases, where flood damage was favored by a majority of participants, are shown depicting the distribution of participants forecasts (Fig 98). The spread for flood damage is much larger than flooding in general. While participant forecasts for these 6 cases compared well with observations for each variable there was still considerable uncertainty with regard to flooding. This disconnect was justified as even on these floodier days it was difficult to pinpoint significant flooding, seemingly because flooding was neither widespread enough nor particularly impactful to high population areas. Thus the challenge is the anticipation of flood verification, i.e. if it doesn't happen where people are then reports tend to be minimal, especially in the overnight hours.

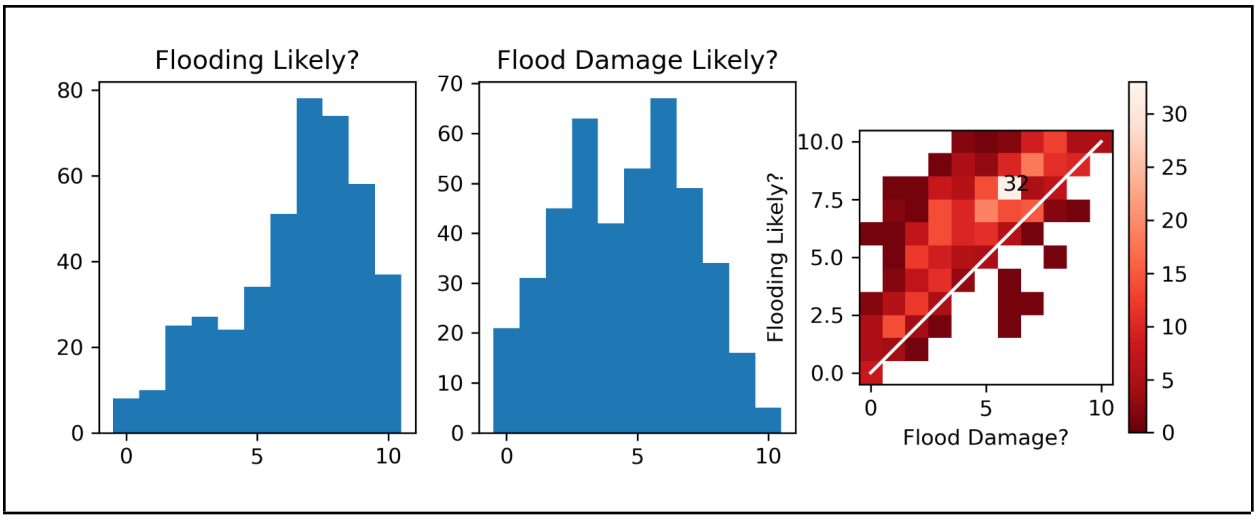


Figure 97: *Histograms of the responses to survey questions regarding flooding as likely (left), flood damage as likely (middle), and the heatmap of paired responses (right). The white line represents the 1:1 line, and 32 is the maximum value in the heatmap.*

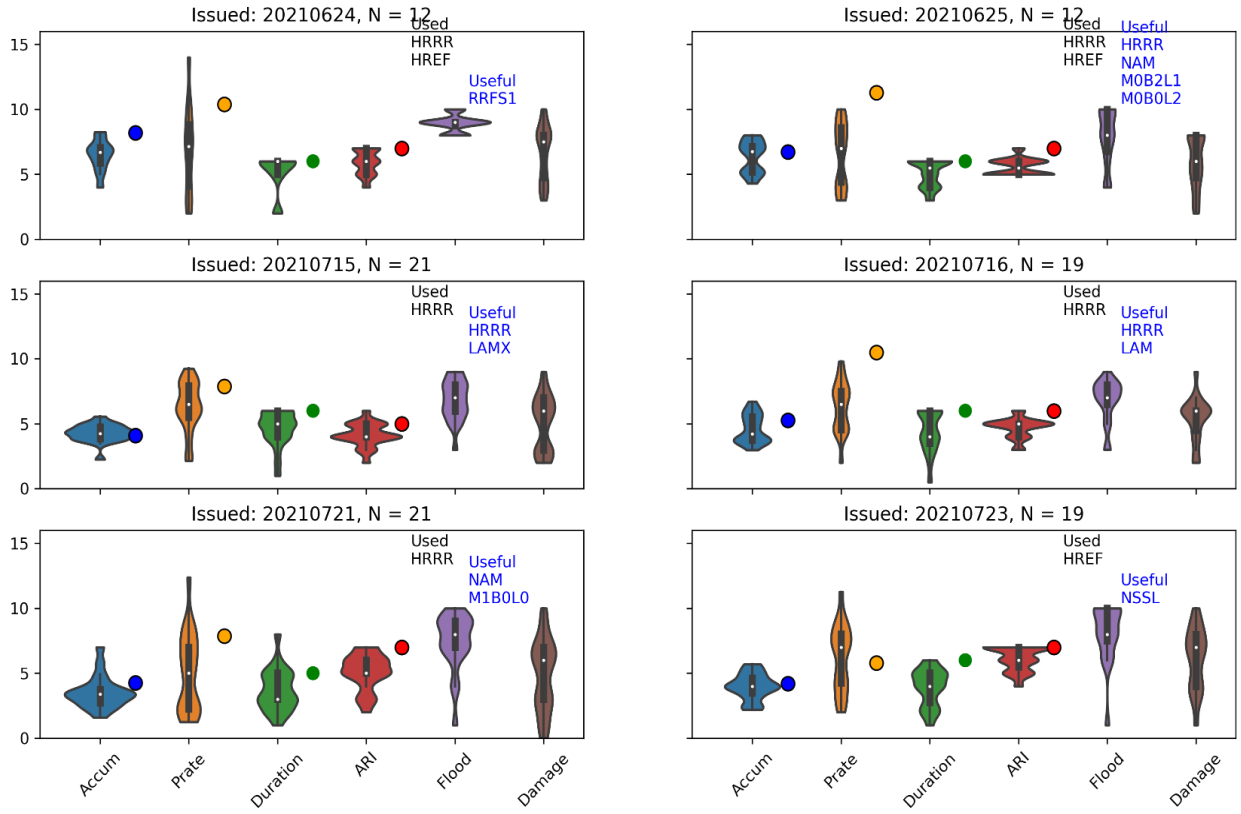


Figure 98: *Sequence of violin plots depicting the participants' forecasts of domain maximum rainfall (blue violin), maximum precipitation rate (orange violin), duration of precipitation rate above 0.25"/hr (green violin), ARI (red violin), flood potential (purple violin), and flood damage potential (brown violin). Dates on top of each chart reflect the issuance day of the forecasts with the number of forecasters in each sample indicated by N, along with annotations of the most used guidance (black), and secondarily the guidance found useful (blue). The respective colored dots represent the verification from MRMS derived information.*

5. Summary and Conclusions

The model data evaluated during the 2021 FFaIR Experiment was centered around numerous configurations of the FV3 on the convective allowing scale: three versions from EMC (referred to as LAMs), one version from GSL (referred to as RRFS1), and four members from the CAPS ensemble (refer to as SSEF members). Also evaluated were two FV3-based ensembles, the SSEF from CAPS and an ensemble that was run in the cloud as a collaboration between GSL, EMC, and NSSL (referred to as RRFSCE). In addition to the models/ensembles evaluated, six versions of the CSU "First Guess" CAM-based Day 1 EROs were examined. This included four versions trained on the NSSL model (see Table 4 for the differences between them), one trained on the HRRR, and one that was a blend of the NSSL2, HRRR, and GEFS trained ERO products, referred to as the BLEND.

The main findings and recommendations are summed up in the following bullet points. Table 7 identifies what the transition recommends are for the guidance.

- EMC provided two models, the **LAM** and **LAMX**, that were identical aside from the domain they were run on. The LAM was run on a domain similar to the HRRR CONUS domain while the LAMX was run on the RRFS North American domain. The goal of this was to determine if the larger domain had a significant impact on the forecast. Both subjective and objective evaluation of the LAM and LAMX show that there was little difference in the QPF forecasts between these two models. **Therefore it is recommended that EMC move forward with running their LAMs on the larger, Northern American grid.**

- EMC also provided a FV3 configuration that included data assimilation, **LAMDAX**. Subjectively this was liked less than the LAM and LAMX by the participants. However, contingency table metrics suggest its performance, at least for the 00z run, is similar to the other two LAMs, with a slightly lower wet bias. **It is recommended that the data assimilation methodology for the LAMs continue to be developed.**

- The **RRFS1** provided by GSL was the least liked FV3- CAM by the participants. Participants often noted that the QPF footprint/storm mode did not resemble observations. Its 00z cycle was the lowest performing FV3-CAM evaluated during FFaIR at both the half inch and one inch 24h QPF thresholds. The 12z cycle's performance was more comparable to the LAM and LAMX, though the RRFS1's wet bias was greater than the LAM/LAMX's wet bias for this cycle. **Recommended for continued development.**

- At the half inch and one inch 24 h QPF threshold, the FV3-CAMs have a similar bias to the operational models, with the SSEF members, RRFS1 and LAMDAX all having a slight dry bias for the half inch threshold. At the higher end 24h QPF thresholds (2+ inches) all FV3-CAMs have a wet bias greater than the HRRR. At two inches the bias is similar to the NAMnest but at three inches the wet bias seen in the FV3-CAMs is greater than the NAMnest wet bias. In some instances the wet bias approached 5 from some models. This suggests that the FV3 models generally underforecast the occurrence of rainfall (dry bias at low thresholds) but when they do forecast rainfall they overforecast the magnitude of the precipitation. **Therefore it is recommended that developers work to identify what is driving the FV3-CAM's difficulty initiating precipitation but once initiating it over forecasting amounts. .**

- The wet bias in the LAMs, RRFS1, and to a lesser extent the SSEF members is very evident when single cell (popcorn) convection is forecasted. The size and shape of the cells resemble grid cells and reminds of the team of the grid point storms that were seen when CAMs were originally being tested. In many instances, especially for the RRFS1,

when the popcorn convection was forecasted nearly every cell had hourly QPF exceeding 2 inches. For the RRFS1 specifically, there were times when the hourly QPF from these cells exceeded 9 inches. **Until identification of what is driving the overproduction of rainfall in single cell convection in the FV3-CAMs, implementation of any of the configurations should not occur.**

- The coverage of instantaneous precipitation rates from the models evaluated during subjective verification was smaller when compared MRMS. However when focusing on the maximum p-prate, the SSEF CNTL member, LAMDAX and RRFS1 generally had a greater maximum than the HRRR or MRMS did. The maximum p-rate from the SSEF CNTL were overall less extreme than they were from the LAMDAX and RRFS1. When evaluating the maximum p-rate from June to July, the average maximum p-rate from the LAMs were roughly double the average maximum from the MRMS and HRRR. The average maximum p-rate from the RRFS1 was 3 to 4 times as great, with some maximums exceeding 150 in/hr. **The FV3-CAMs, especially the RRFS1, have a tendency to output p-rates that are much larger than observed and at times that are unrealistic. It is possible that these rates are helping to drive the large QPF values seen in the single cell convection from the models. The impact of these p-rates and what is driving such high magnitudes should be examined further.**

- Neither the SSEF or the RRFSCE outperformed the HREF. When focusing on probability of exceedance thresholds, the SSEF probabilities were extremely low. The RRFSCE probabilities were felt to be comparable to the HREF probabilities at times by the participants, but this was generally when the ensemble appeared to have a good handle on the pattern. When this occurred, probabilities were generally high due to the low spread in the ensemble. The mean products (traditional mean, PMM, and LPMM) from the SSEF, for the most part, had a slightly higher CSI than the respective means of the RRFSCE. The RRFSCE means had a similar bias to the HREF while the SEFF means consistently had a lower bias than the RRFSCE and HREF. Interestingly, the PMM, which is known for having a high bias, from the SSEF at a half inch and an inch had a dry bias. Therefore, it is likely that the SSEF overall is less likely to forecast precipitation in general. **It is recommended that both the SSEF and RRFSCE should continue with development. The RRFSCE appears to suffer from underdispersion and ways to introduce greater spread into the model should be examined.**

- All of the NSSL-based CSU “first-guess” EROs, except NSSL5, performed similarly to one another. The NSSL2 was the favorite of the participants while NSSL4 had a slight edge over the other ERO models in the statistical analysis. Based on frequency of issuance, the NSSL4 is more likely to forecast higher ERO risk categories than the other models and the FFaIR ERO. However, this is not necessarily a negative as participants and WPC forecasters have commented that they prefer the “first guess” to have a slight high bias since it alerts them to where they need to focus their analysis on. **It is**

recommended that CSU continues development on all of the NSSL EROs, especially the configurations for NSSL2 and NSSL4.

- The HRRR-based ERO had the lowest performance during FFaIR, both subjectively and objectively. However this was likely in part due to the Monsoon being the dominant forecast change during the last two weeks of FFaIR. As discussed in Section 4.3.1, the HRRR model QPF had a dry bias across the southwest during the Monsoon. Numerous things such as lack of observations across Mexico to ingest into the HRRR's DA help lead to the low bias across the region. This likely impacted the performance of the HRRR-based ERO model across the southwest. Additionally, the HRRR-based ERO has a short training period that does not include an active Monsoon season. **Therefore it is recommended that the HRRR-based ERO training period is adjusted to include this year's warm season, and thus the Monsoon.**

- The CSU BLEND ERO was the most liked "first guess" ERO by the participants during the first half of FFaIR. During the second half, its performance was hampered by the poor performance of the HRRR-based ERO, which is one of the three ERO models used to create the BLEND ERO. **The BLEND ERO should continue to be refined and perhaps re-evaluation of how the weights of each ERO model are determined in the creation of the BLEND ERO.**

Table 7:*Research to Operations Transition Metrics for the 2021 FFaIR Experiment.*

Models, Ensembles and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
LAMX		X		EMC
LAMDAX		X		
RRFS1		X		GSL
RRFSCE		X		EMC/GSL/NSSL
SSEF		X		OU/CAPS
CSU-ML Day 1 ERO NSSL2/NSSL3/ NSSL4/NSSL5		X		CSU/JTTI
CSU-ML Day 1 ERO HRRR		X		
CSU-ML Day 1 ERO BLEND		X		

Acknowledgments

The FFaIR team would like to extend an enormous thanks to everyone who helped us prepare for the experiment, those who provided guidance, those who participated in FFaIR (Appendix A.1) and all those who gave a science seminar (Appendix A.2). A special thanks to **Ben Albright** and **Mike Erickson** for helping with the verification of the guidance evaluated and for their help editing the Final Repor. To our WPC Forecasters (**Marc Chenard, Josh Weiss, Bryan Jackson,** and **Greg Gallina**), we can not thank you enough for all your hard work and for leading the forecast discussions during the experiment. Your insight and guidance truly helped enhance the participants' experience and you provided great insight into the difficulties of forecasting the heavy rainfall/flash flooding risk for the whole CONUS. Again, thank you to those who provided data for FFaIR: GSL, EMC, CAPS, CSU, and Keith Brill. We enjoy working with you and look forward to our continued partnership. **Diana Stovern** thank you for your help in creating the ARI exceedance plots of model QPF for the FFaIR website. Lastly, the team would also like to send a huge thank you to our intern **Miranda Bitting**. Your hard work during FFaIR and after FFaIR evaluating p-rate information was extremely helpful. We are very proud of you and are amazed by your ability to rise to any task we handed you, including running the ERO/ARI-ERO breakout groups.

References

ABC7 Staff, 2021: Arizona flash flooding prompts dramatic rescue of man, 2 daughters. Accessed July 15, 2021, https://abc7.com/arizona-flooding-flash-water-rescue-catalina/10888440/

Blumhardt, M., 2021: Meteorologist: Here's why the Poudre River flooded and why it likely will again this year. Accessed July 23, 2021, https://www.coloradoan.com/story/news/2021/07/23/poudre-river-flooding-explained-colorado-meteorologist-black-hollow-flood/8063039002/

Bradford, C., 2021: Arizona floods: Monsoon hits Flagstaff with cars washed down streets & terrified residents being warned to lockdown. Accessed July 16, 2021, https://www.the-sun.com/news/3285835/arizona-floods-monsoon-flagstaff-cars-washed-streets-lockdown/

Cappucci, M., 2021: Eight inches in one hour: How a deadly downpour flooded Zhengzhou, China. Accessed July 21, 2021, https://www.washingtonpost.com/weather/2021/07/21/zhengzhou-china-record-rain-flooding/

Crowe, K.C. II and P. DeMola, 2021: Major damage from heavy rains in Rensselaer County. Accessed July 15, 2021, https://www.timesunion.com/news/article/Heavy-rains-flooding-close-roads-in-Rensselaer-16315342.php#photo-21239170

Erickson, M. J., J. S. Kastman,, B. Albright, S. Perfater, J.A. Nelson, R.S. Schumacher, and G.R. Herman, 2019: Verification Results from the 2017 HMT–WPC Flash Flood and Intense Rainfall Experiment. *J. Appl. Meteor. Climatol.*, **58**, 2591-2604, https://doi.org/10.1175/JAMC-D-19-0097.1

Erickson, M. J., B., Albright, and J. A. Nelson, 2021: Verifying and Redefining the Weather Prediction Center's Excessive Rainfall Outlook Forecast Product. *Wea. Forecasting*, **36**, 325-340, https://doi.org/10.1175/WAF-D-20-0020.1

FOX 13 News, 2021: Rain and rising floodwaters play role in southern Utah train derailment. Accessed July 16, 2021, https://www.fox13now.com/news/local-news/rain-and-rising-floodwaters-play-role-in-southern-utah-train-derailment

Gan, N. and Z. Wang, 2021: Death toll rises as passengers recount horror of China subway floods. Accessed August 10, 2021, https://www.cnn.com/2021/07/22/china/zhengzhou-henan-china-flooding-update-intl-hnk/index.html

Kottasová, I. and M. Krever, 2021: Deadly floods inundated parts of Europe, but the Netherlands avoided fatalities. Here's why. Accessed on July 21, 2021, https://www.cnn.com/2021/07/19/world/netherlands-germany-flood-defense-warning-system-intl-cmd/index.html

LaPointe, S., 2021: Monday July 19, 2021: Isolated Fonda Flash Flood Event. Accessed on July 20, 2021, http://slapointewx.com/wrgb/weather_historical_daily/2021/Fonda_Flood.html

Loveday, M., J. Hassan, and L. Beck, 2021: At least 95 dead and some 1,300 missing as flooding rages across Europe. Accessed July 16, 2021, https://www.washingtonpost.com/world/2021/07/15/germany-flooding-buildings-collapse/

Marthiesen, K., H. Burchard and L. Gehrke, 2021: Over 100 die in Germany, Belgium floods despite early warnings. Accessed July 21, 2021 https://www.politico.eu/article/germany-floods-dozens-dead-despite-early-warnings/

Meckles, J. and C. Bianchi, 2021: I-70 Reopens in Glenwood Canyon. Accessed July 22, 2021, https://www.9news.com/article/weather/weather-colorado/i-70-closed-at-avon-and-glenwood-canyon-due-to-flash-flooding/73-06194b6e-9721-4e41-bb66-59153df1d1cc

Puca, S., L. Brocca, Panegrossi, G., Mascitelli, A., Krahe P., Fairbairn D., Baguis, P., Zauli F., Mahovic, N. S., Chatterjee, C. T., and Smilijanic, I., 2021: Devastating floods in western Europe. Accessed from EMUMETSAT website on July 21, 2021, https://www.eumetsat.int/devastating-floods-western-europe

Reed, C., 2021:`Worst I've ever seen`: Flash Flood Spreads Mud Through Springdale, Closes Zion. Accessed June 30, 2021, https://www.stgeorgeutah.com/news/archive/2021/06/29/cdr-worst-ive-ever-seen-flash-flood-spreads-mud-through-springdale-closes-zion/#.YTt4PhmSlaS

Sarles, J., 2021:Some Streets Flood In Greeley After 3 To 4 Inches Of Rain Falls In An Hour. Accessed July 1, 2021, https://denver.cbslocal.com/2021/07/01/street-flooding-greeley-heavy-rain-standing-water

Stroud, J., 2021: Rains the Triggered Mudslides in Glenwood Canyon Approached what Weather Observers Call a 500-year Event. Accessed August 5, 2021, https://www.postindependent.com/news/rains-that-triggered-mudslides-in-glenwood-canyon-approached-what-weather-observers-call-a-500-year-event/

The Guardian, 2021: Death toll exceeds 180 as Germany and Belgium hit by devastating floods. Accessed August 6, 2021, https://www.theguardian.com/world/2021/jul/16/western-germany-floods-angela-merkel-horror-catastrophe-deaths-missing-search-flooding-belgium

Trojniak, S.M. and J. Correia, 2021: 2021 Flash Flood and Intense Rainfall (FFaIR) Experiment: Program Overview and Operations Plan. 28 pp, https://docs.google.com/document/d/1aCcQsffKKCcx69YIxAwPThkvUaEvInPIblt9fS88kfM/edit?usp=sharing

Trojniak, S.M. and J. Correia, 2020: 2020 Flash Flood and Intense Rainfall (FFaIR) Experiment: Final Report. 104 pp, https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf

WNYT Staff, 2021: Fonda state of emergency lifted after flooding. Accessed on July 20, 2021, https://wnyt.com/montgomery-county-ny-news/state-of-emergency-flooding-fonda-montgomery-county/6176967/

Zhou, M.: Email about the Extreme rainfall and flooding in Zhengzhou, China sent to the NOAA MAP email listserv.

Appendix A

A.1 List of Participants

Table A.1: *List of the participants for each week of the 2021 FFaIR Experiment. Note the experiment did not run during the week of the Fourth of July.*

Week	WPC Forecaster	WFO/RFC/other	Research/Academia/ Student	EMC and GSL	Other Regional and National Centers/Offices
Week 1 June 21 – 25	Marc Chenard	Bekki Harjo - TUARFC Nick Webb - WFO RLX Kathleen Pelczynski -WFO RAH Evan Webb - WFO LMK	Aaron Hill - CSU Allie Mazurek - CSU	John Brown - GSL Matthew Pyle - EMC Marcel Caron - EMC Geoff Manikin - EMC	Chandra Kondragunta - JTTI Program Manager Heather Grams - NSSL Javier Villegasbravo - WPC
Week 2 June 28 – July 2	Josh Weiss	Matt Anderson - WFO HUN Robert Deal - WFO PHI Eswar Iyer - WFO AKQ Michael Brown - WFO PBZ Keith Cooley - WFO GRB Brent Hewett - WFO MPX		Ed Szoke - GSL Shannon Shields - EMC	Felicia Guarriello - OAR Kate Abshire - WRSB Race Clark - NSSL
Week 3 July 12-16	Bryan Jackson	Charles Ross - WFO CTP Andrew Zimmerman - WFO AKQ Jay Engle - WFO OKX Craig Schmidt - WFO HUN Mike Jamski - WFO CYS Megan Terry - WFO SGF Shawn DeVinny - NCRFC Randy Bowers - WFO OUN Christina Leach - WFO MOB	Russ Schumacher - CSU Bill Gallus - Iowa State Anna Wanless - OU	Eric James - GSL and CSU Ben Blake - EMC Logan Dawson - EMC	Jordan Rabinowitz - ERH/SSD Caleb Steele - WR Headquarters
Week 4 July 19-23	Greg Gallina	Jason Alumbaugh - WFO BUF Matt Steinbugl - WFO CCTP Andrei Evbuoma - WFP ALY TJ Turnage - WFO GRR Sarah Marquardt - WFO MPX Brian Astifan - OHRFC Helge Tuschy - German Weather Service	Keith Brewster - OU CAPS Jacob Escobedo - CSU	Jeff Duda - GSL Anning Cheng - EMC Chris MacIntosh - EMC	Eric Allen - ERH/SSD Julie Lesko - SSH for SR Kent Knopfmeier - NSSL/HWT Javier Villegasbravo - WPC

A.2 List of the Science Seminars

Table A2: *List of the presenters and the title of their presentations during the 2021 FFaIR Experiment. Two presentations were given on Tuesdays and one on Thursdays. The presentations can be found in this shared folder.*

Presenter	Presentation Title
WEEK 1 (June 21-25)	
Aaron Hill (CSU)	CSU CAM-based First Guess Excessive Rainfall Outlook Products
Jacob Carley (EMC)	Clouds in the Cloud: Using Cloud HPC to Develop NOAA's Next Generation High Resolution Forecast System
Jian Zhang (NSSL)	MRMS radar QPE version 12.2
Steve Martinaitis (NSSL)	MRMS Gauge Quality Control and Multi-Sensor QPE
WEEK 2 (June 28-July 2)	
Daphne Ladue and Alex Marmo (OU CAPS)	They Don't Only Care About Severe Weather: Our Accidental Flood Study
Russ Schumacher (CSU)	How Do We Define Excessive Rainfall? (And Why it Matters)
JJ Gourley (NSSL)	Precipitation Proxies and MRMS
WEEK 3 (July 12-16)	
Theodora Meredith (WFO American Samoa)	Forecasting heavy rainfall in American Samoa with No Radar!
Bill Gallus	Accounting for Spatial Displacement Errors in Ensemble Member QPF to Create Short-Term Ensemble Streamflow Forecasts
Eric James (GSL/CSU)	Comparing QPF and QPE vs. Precipitation Thresholds for Flash Flood Forecasting and Aanalysis
Anna Wanless (OU Grad Student)	Examining extreme Rainfall Forecast and Communication Processes in the South Central United States
WEEK 4 (June 19-23)	
Kelly Mahoney (PSL)	Unscrambling Omelets? Disentangling Errors to Improve Ensemble Hydrometeorological Predictions
Alex Lamers (WPC)	Possibilities for the Future of the WPC Met Watch Desk and the Watch-Warning Gap
Keith Brewster (OU CAPS)	CAPS Storm Scale Ensemble Forecast (SSEF) for the HMT FFaIR Experiment
Jordan Dale (OAR WPO)	An Overview of OAR's Weather Program Office (WPO)

Appendix B

Guidance and Products Evaluated

Table B.1 *The deterministic and ensemble model guidance and products that will be evaluated in the 2021 FFaIR experiment. Light blue indicates operational guidance, dark blue indicates experimental guidance.*

Provider	*Model*	*Resolution*	*Forecast Hours*	*Notes*
ESRL/GSL	***HRRR***	*3 km*	*Hourly forecasts. Forecast length: 00, 06, 12, and 18 UTC runs are 48 h. All other run times are 18 h, as is the sub-hourly output.*	*High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model.*
EMC	***HREFv3***	*~3 km*	*48 h forecast run daily at 00 and 12 UTC.*	*Consists of 10 members, each member provides a real-time and time-lagged run.*
EMC/WPC/ MDL	***PQPF from GEFS, WPC, and NMB***		*PQPF out to Day 3 at 00, 06, 12, and 18 UTC*	*6 h and 24 h QPF for the 90th, 95th, and 90th percentiles.*
GFDL/EMC	***FV3-LAM*** *(Limited Area Model)*	*~3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*The LAM is the stand-alone regional version of FV3 run over the CONUS domain. Configuration uses the Thompson microphysics and MYNN planetary boundary layer schemes.*

GFDL/EMC	***FV3-LAMX***	*~3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*Same configuration as FV3-LAM but run on the North American grid.*
GFDL/EMC	***FV3-LAMDAX***	*~3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*Experimental Stand-alone Regional with hourly data assimilation from the RAP, same physics suite as FV3-LAM.*
ESRL/GSL	***RRFS-dev 1***	*~3 km*	*Hourly forecasts. Forecast length: 00, 06, 12, and 18 UTC runs out to 60 h. All other run times out 18 h*	*HRRRv4-like over CONUS. Hourly cycled DA from GDAS ensemble and IC/LBC[46] from a 13km North American FV3-LAM.*
ESRL/GSL	***RRFS-dev 2***	*~3 km*	*00 and 12 UTC out 60 h..*	*HRRRv4-like over North American domain. GFS for ICs/LBCs. No DA.*
EMC/GSL/NSSL	***RRFS[47]-Cloud***	*~3 km*	*00 UTC run out to 60 h*	*North American Domain. Uses three physics suites, For each physics suite: one member not perturbed while the other 2 members are perturbed, but using different methods.*

[46] DA: Data Assimilation
IC: Initial Conditions
LBC: Lateral Boundary Conditions
[47] RRFS: Rapid Refresh Forecast System

OU/CAPS	***SSEF***	*3 km*	*00 UTC run out to 84 h*	*CONUS only. 15 member ensemble made from mixed physics.*
OU/CAPS	***SSEF CNTL Member (M0B0L0)***	*3 km*	*00 UTC run out to 84 h*	*Microphysics: Thompson [48]LSM: NOAH*
OU/CAPS	***SSEF RRFS CNTL Member (M0B0L1)***	*3 km*	*00 UTC run out to 84 h*	*Microphysics: Thompson LSM: NOAH-MP*
OU/CAPS	***SSEF HRRR-like Member (M0B0L2)***	*3 km*	*00 UTC run out to 84 h*	*Microphysics: Thompson LSM: RUC*
OU/CAPS	***SSEF WoFS-like Member (M1B0L0)***	*3 km*	*00 UTC run out to 84 h*	*Microphysics: NSSL LSM: NOAH*

[48] LSM: Land-surface model

Appendix C

MRTP Survey

What is your point forecast of Maximum Rainfall Accumulation inside your contour (inches, decimal number only, i.e. 1.01) ? *

Your answer

What is your forecast of Maximum Rainfall Rate (inches per hour, decimal number only; i.e. 2.75)? *

Your answer

What is your forecast of the point maximum 6h ARI? *

- ◯ 1 Year
- ◯ 2 Year
- ◯ 5 Year
- ◯ 10 Year
- ◯ 25 Year
- ◯ 50 Year
- ◯ 100 Year
- ◯ None

Could this rainfall result in flooding? *

	0	1	2	3	4	5	6	7	8	9	10	
No flooding	○	○	○	○	○	○	○	○	○	○	○	Flooding

How much damage might the flooding cause? *

	0	1	2	3	4	5	6	7	8	9	10	
No Damage	○	○	○	○	○	○	○	○	○	○	○	Damaging

What will be the maximum duration (integer, whole hours) of the rainfall rate exceeding 0.25"/hr? *

Your answer

Which model were you assigned to use? *

- ◯ EMC LAM
- ◯ EMC LAMX
- ◯ EMC LAMDAX
- ◯ NSSL
- ◯ GSL RRFS 1
- ◯ SSEF CNTL Member
- ◯ SSEF RRFS-like Member
- ◯ SSEF HRRR-like Member
- ◯ SSEF WoFS-like Member
- ◯ HRRRv4
- ◯ NAMnest
- ◯ GFSv16
- ◯ None
- ◯ Other: _____

Which ensemble were you assigned to use? *

- ◯ GEFS
- ◯ HREFv3
- ◯ RRFS Cloud
- ◯ CAPS Ensemble - SSEF
- ◯ None
- ◯ Other: _____

Which models/ensembles did you strongly consider in your final forecast *

	Used	Useful	Considered	Not considered	Not available
EMC LAM	☐	☐	☐	☐	☐
EMC LAMX	☐	☐	☐	☐	☐
EMC LAMDAX	☐	☐	☐	☐	☐
NSSL	☐	☐	☐	☐	☐
RRFS 1	☐	☐	☐	☐	☐
SSEF CTL Member	☐	☐	☐	☐	☐
SSEF WoFS-like Member	☐	☐	☐	☐	☐
SSEF RRFS-like Member	☐	☐	☐	☐	☐
SSEF HRRR-like Member	☐	☐	☐	☐	☐
HRRRv4	☐	☐	☐	☐	☐
NAMnest	☐	☐	☐	☐	☐
GFSv16	☐	☐	☐	☐	☐
GEFS	☐	☐	☐	☐	☐
HREFV3	☐	☐	☐	☐	☐
GSL Cloud	☐	☐	☐	☐	☐
CAPS Ensemble - SSEF	☐	☐	☐	☐	☐

The phenomenon you are forecasting for involves: *

- [ ] Mesoscale Convective System
- [ ] Supercells
- [ ] Multi-cell clusters
- [ ] Tropical Cyclone
- [ ] Pulse Thunderstorms
- [ ] Training convective elements
- [ ] fast moving convection
- [ ] Popcorn convection
- [ ] Stratiform rainfall
- [ ] Other:

Please discuss the utility or usefulness of your assigned model/ensemble. Include primary concerns such as magnitude, timing, phenomena that may have contributed to your assessment.

Your answer